Renewable Energy: A Case Study of Wind Power Plant

Yifei Chen (cheny70@rpi.edu), Danqi Jiang (jiangd4@rpi.edu), Wanjing Song (songw5@rpi.edu), Weijun Li(liw18@rpi.edu), Xin Ning(ningx4@rpi.edu), Xiao Zou(zoux2@rpi.edu)

[1]Rensselaer Polytechnic Institute, Tetherless World Constellation, Troy, NY, United States,

Abstract

Wind power, which is one of the most popularly used renewable energy sources all over the world, is sensitive to change of wind speed under various climate conditions. According to the relationship between the wind plant power generation and local wind speed, the impact of wind speed can be analyzed. In this study, we merged power plant data with NASA wind speed data and conducted some exploratory data analysis by making a histogram, heatmap and boxplot for the preprocessed dataset. We developed linear and nonlinear regression models. The result shows that the regression performed poorly on the dataset. K-means help us get the relationship about the number of groups and squared error. 7 is the best group divide with the silhouette score around 0.9. After power generation is computed into seven groups based on K-means, then two classification models are implemented to show the major influence for power generation. Both Random Forest Classifier and XGBClassifier conclude that the amount of natural gas used in power plants contribute most to generation capacity among features of other conventional energy sources (oil, coal), wind speed, and geographic locations. Although the accuracy of each model is over 90%, both models have overfitting issues. All the models show that the related energy used, speed and location are important features that influence the power output of a power plant. In the future, our study may give people more insights about how to locate a power plant to a better location and improve the production efficiency.

Data Description

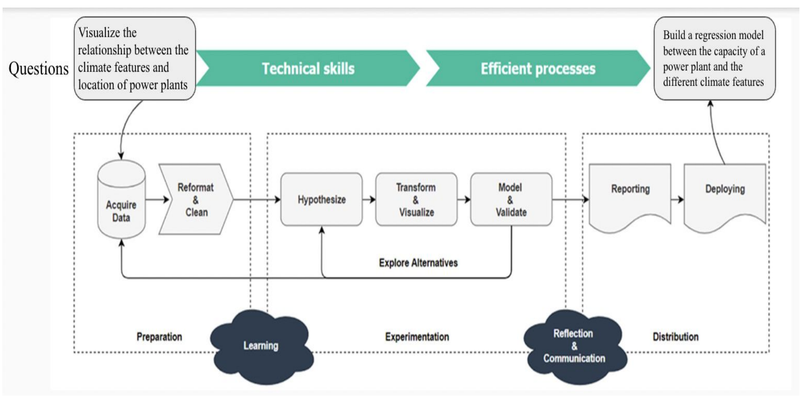
Problem Area

- Determine the relationship between the wind speed and power plant which used wind energy
- Analyze the effect of wind speed and other related features on the location/capacity of a power plant

Datasets

- **M2T1NXFLX (or tavg1_2d_flx_Nx)** from the NASA website which includes surface air temperature, surface specific humidity, surface wind speed, etc.
- **Power plant** dataset it includes power plants from all over the world and the plant types include hydroelectric dams, fossil fuel (coal, natural gas, or oil), nuclear, solar, wind, geothermal, and biomass.

Workflow



Our project workflow has three major steps: Acquire data, build the model for analysis, and make the conclusion. First, we acquire wind speed data from NASA and power plants data from an energy website. Then we can convert the netcdf file . Then we will make hypotheses such as the location of the plant with the power generation. Next, we used python to build the model such as linear regression,K-means and so on to analyze the relationship between different factors. After we build the model to examine whether our hypothesis is correct or not, it is important to help us to make the conclusion and examine whether our hypothesis is correct or not. The last step is to make the presentation and then make the report of our job and our model to the class and show our job to the public.

https://github.com/ITWSDataScience/Group4_2022 github

Sponsors:

Tidying Messy Datasets

1. Import data:

Python (P)

A. Spreadsheet files

P:

First dataset :GPM Level 3 IMERG *Final* Daily 10 x 10 km (GPM_3IMERGDF)

```
data = Dataset('/content/drive/MyDrive/Data Science/nasa.nc4', mode='r')
```

Second Dataset:power plant dataset

```
df2 = pd.read_csv('/content/drive/MyDrive/Data Science/Power_Plants.csv')
```

Merge the two dataset

```
final_df = pd.merge(df,df2_transform, left_on=['lon','lat'], right_on = ['LONGITUDE','LATITUDE'], how='right')
```

2. Inspect data:

Several methods are used to inspect the dataset.

A. Get an overview of the dataframe (final_df):

P:

```
final_df
```

	lon	lat	speed	LONGITUDE	LATITUDE	OPER_CAP	COAL_USED	NGAS_USED	OIL_USED
0	-122.500	41.5	4.161290	-122.500	41.5	26.0	0.0	0.0	0.0
1	-122.500	41.5	4.161290	-122.500	41.5	36.0	0.0	0.0	0.0
2	-122.500	41.5	4.161290	-122.500	41.5	19.0	0.0	0.0	0.0
3	-113.125	38.0	5.384007	-113.125	38.0	44.8	0.0	0.0	0.0
4	-111.875	44.0	6.046074	-111.875	44.0	6.7	0.0	0.0	0.0

B. Check the missing values in the dataset

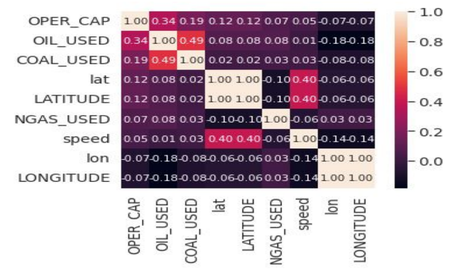
```
final_df.isnull().sum()
```

```
lon          2314
lat          2314
speed        2314
LONGITUDE    2314
LATITUDE     2314
OPER_CAP     2314
COAL_USED    2314
NGAS_USED    2314
OIL_USED     2314
dtype: int64
```

C. Exploratory Data Analysis

Three commonly used exploratory plots are heatmap, box, and histograms.

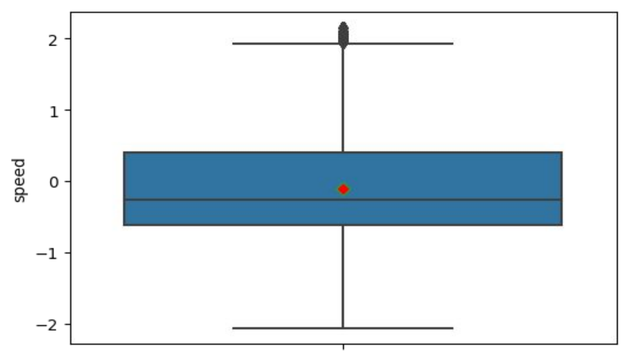
i. Heatmap-1



ii. Box plot

P:

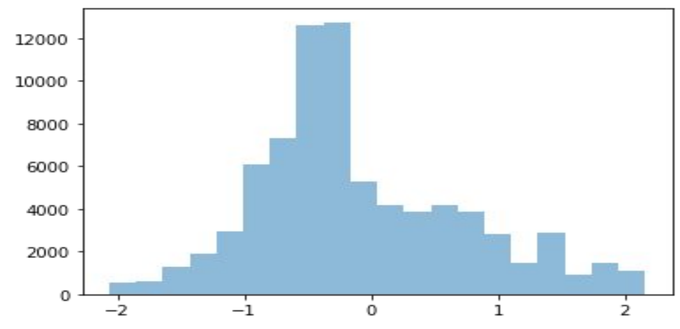
```
sns.boxplot(y=df2,showmeans=True,meanprops = {'marker':'D','markerfacecolor':'red'},)
```



iii. Histogram plot

P:

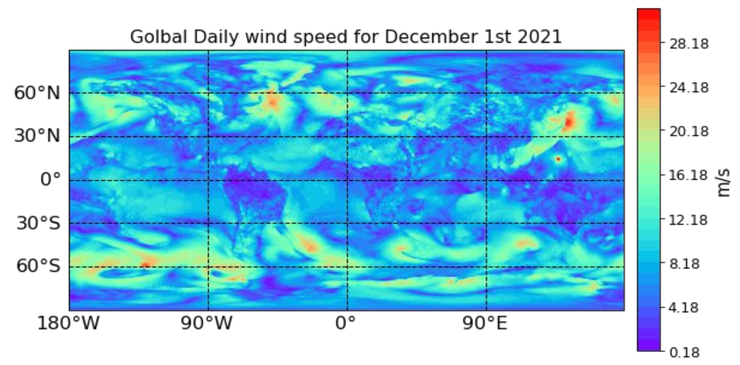
```
plt.hist(df2, bins=20,  alpha=0.5)
```

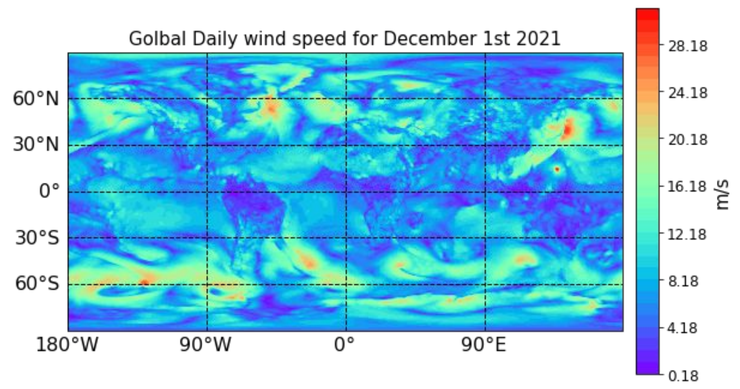


3.Model:

A. -Regression model

Visualize the geospatial dataset on the map





Linear Regression model

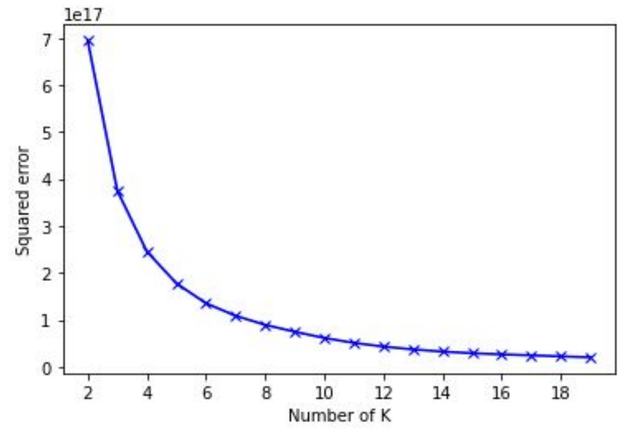
R-square of linear regression	MSE of linear regression
0.0088	854.26

SVR with rbf model

R-square of svr	MSE of linear svr
0.09	942.5

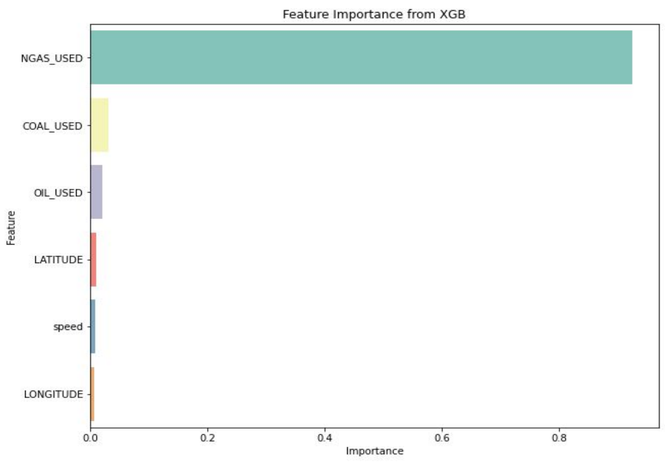
B. K-means

K-means to get group for classification



C. Classification

Feature importance from Classification



Glossary:

Python – A programming language, capable of processing data/statistical analysis

Resources:

Datasets:

https://disc.gsfc.nasa.gov/datasets/M2T1NXFLX_5.12.4/summary?keywords=wind%20speed&start=2021-12-01&end=2022-12-01

https://hifld-geoplatform.opendata.arcgis.com/datasets/geoplatform::power-plants-2/about?layer=0

Reference:

Burak. *Location and Investment Factors of Hydropower Plants*. https://www.tandfonline.com/doi/full/10.1080/15567036.2021.1963015.

De Souza Dias, Viviane, et al. “An Overview of Hydropower Reservoirs in Brazil: Current Situation, Future Perspectives and Impacts of Climate Change.” *MDPI*, Multidisciplinary Digital Publishing Institute, 3 May 2018, https://www.mdpi.com/2073-4441/10/5/592.