**Problem Statement**

The Global Power Plant Database is a comprehensive, open source database of power plants around the world. It centralizes power plant data to make it easier to navigate, compare and draw insights for one’s own analysis. The database covers approximately 35,000 power plants from 167 countries and includes thermal plants (e.g. coal, gas, oil, nuclear, biomass, waste, geothermal) and renewables (e.g. hydro, wind, solar). Each power plant is geolocated and entries contain information on plant capacity, generation, ownership, and fuel type. It will be continuously updated as data becomes available.

Prediction : Make two prediction

1) Primary Fuel

2) capacity\_mw

**Data Analysis**

* **There are 207 rows and 27 columns**
* **The features in the above data set are-**
* latitude (geolocation in decimal degrees)
* longitude (geolocation in decimal degrees)
* commissioning\_year (year of plant operation, weighted by unit-capacity when data is available)
* geolocation\_source (attribution for geolocation information)
* other\_fuel1 (energy source used in electricity generation or export)
* generation\_gwh\_2014 (electricity generation in gigawatt-hours reported for the year 2014)
* generation\_gwh\_2015 (electricity generation in gigawatt-hours reported for the year 2015)
* generation\_gwh\_2016 (electricity generation in gigawatt-hours reported for the year 2016)
* generation\_gwh\_2017 (electricity generation in gigawatt-hours reported for the year 2017)
* generation\_gwh\_2018 (electricity generation in gigawatt-hours reported for the year 2018)
* **The label in the given dataset is-**
* primary\_fuel [Classification] (energy source used in primary electricity generation or export)
* capacity\_mw [Regression] (electrical generating capacity in megawatts)
* **The columns not required-**
* country (3 character country code corresponding to the ISO 3166-1 alpha-3 specification)

- *As the entire column has only India as the country, it has no influence on the model*

* country\_long (longer form of the country designation)

- *As the entire column has only India as the country, it has no influence on the model*

* name (name or title of the power plant, generally in Romanized form)

- *As each name is unique, it does not influence the model*

* gppd\_idnr (10 or 12 character identifier for the power plant)

- *Each plant has its own unique id, not influencing the model*

* owner (majority shareholder of the power plant, generally in Romanized form)

*- As the owner of each plant is almost unique, hence not influencing the model*

* source (entity reporting the data; could be an organization, report, or document, generally in Romanized form)

*- As the source of each plant is almost unique, hence not influencing the model*

* url (web document corresponding to the source field)

*- Each plant has its own url, not influencing the model*

* wepp\_id (a reference to a unique plant identifier in the widely-used PLATTS-WEPP database.)

*- Each plant has its own unique id, not influencing the model*

* other\_fuel2 (energy source used in electricity generation or export)

*- As the column is empty it is safe to drop it*

* other\_fuel3 (energy source used in electricity generation or export)

*- As the column is empty it is safe to drop it*

* year\_of\_capacity\_data (year the capacity information was reported)

*- As the year is 2019 as same for all, will not influence the model*

* generation\_gwh\_2013 (electricity generation in gigawatt-hours reported for the year 2013)

*- As the column is empty it is safe to drop it*

* generation\_gwh\_2019 (electricity generation in gigawatt-hours reported for the year 2019)

*- As the column is empty it is safe to drop it*

* generation\_data\_source (attribution for the reported generation information)

*- As the entire column is filled with one value "Central Electricity Authority", it has no significant influence on the model*

* estimated\_generation\_gwh (estimated electricity generation in gigawatt-hours)

*- As the column is empty it is safe to drop it*

* **The data types of features and labels are as follows-**

|  |  |
| --- | --- |
| **Features** | **Data type** |
| latitude | float64 |
| longitude | float64 |
| other\_fuel1 | object |
| commissioning\_year | float64 |
| geolocation\_source | object |
| generation\_gwh\_2014 | float64 |
| generation\_gwh\_2015 | float64 |
| generation\_gwh\_2016 | float64 |
| generation\_gwh\_2017 | float64 |
| generation\_gwh\_2018 | float64 |
|  | |
| **Labels** | **Data type** |
| capacity\_mw | float64 |
| primary\_fuel | object |

* **The null values in features and columns are as follows-**

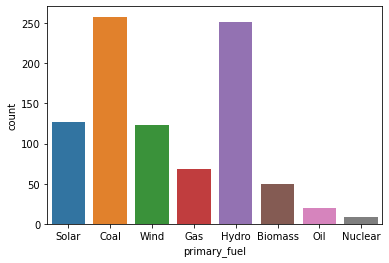
|  |  |
| --- | --- |
| **Features** | **Null value** |
| latitude | 46 |
| longitude | 46 |
| other\_fuel1 | 709 |
| commissioning\_year | 380 |
| geolocation\_source | 19 |
| generation\_gwh\_2014 | 509 |
| generation\_gwh\_2015 | 485 |
| generation\_gwh\_2016 | 473 |
| generation\_gwh\_2017 | 467 |
| generation\_gwh\_2018 | 459 |
|  | |
| **Labels** | **Null value** |
| capacity\_mw | 0 |
| primary\_fuel | 0 |

* **Statistical Analysis of the data-**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **latitude** | **longitude** | **other\_fuel1** | **commissioning\_year** | **geolocation\_source** | **generation\_gwh\_2014** | **generation\_gwh\_2015** | **generation\_gwh\_2016** | **generation\_gwh\_2017** | **generation\_gwh\_2017** | **capacity\_mw** | **primary\_fuel** |
| **count** | 861.000000 | 861.000000 | 907.000000 | 527.000000 | 907.000000 | 398.000000 | 422.000000 | 434.000000 | 440.000000 | 448.000000 | 907.000000 | 907.000000 |
| **mean** | 21.197918 | 77.464907 | 2.777288 | 1997.091082 | 1.754135 | 2431.823590 | 2428.226946 | 2467.936859 | 2547.759305 | 2600.804099 | 326.223755 | 3.206174 |
| **std** | 6.239612 | 4.939316 | 0.429348 | 17.082868 | 0.700183 | 4026.440035 | 4194.596959 | 4162.884308 | 4196.991169 | 4314.880456 | 590.085456 | 2.280652 |
| **min** | 8.168900 | 68.644700 | 0.000000 | 1927.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| **25%** | 16.773900 | 74.256200 | 3.000000 | 1988.000000 | 2.000000 | 223.557672 | 176.381063 | 188.285252 | 177.874930 | 193.378250 | 16.725000 | 1.000000 |
| **50%** | 21.780000 | 76.719500 | 3.000000 | 2001.000000 | 2.000000 | 801.123775 | 711.181225 | 737.205450 | 817.977250 | 751.644375 | 59.200000 | 3.000000 |
| **75%** | 25.512400 | 79.440800 | 3.000000 | 2012.000000 | 2.000000 | 3035.306250 | 3084.121250 | 3282.861313 | 3275.690475 | 3143.535900 | 385.250000 | 6.000000 |
| **max** | 34.649000 | 95.408000 | 3.000000 | 2018.000000 | 3.000000 | 28127.000000 | 30539.000000 | 30015.000000 | 35116.000000 | 35136.000000 | 4760.000000 | 7.000000 |

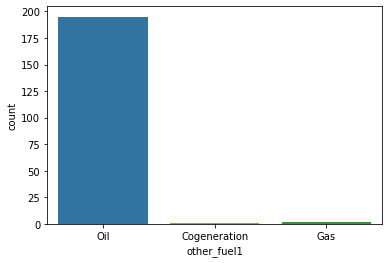
* **Graphical Analysis of Data-**

**primary\_fuel**

****

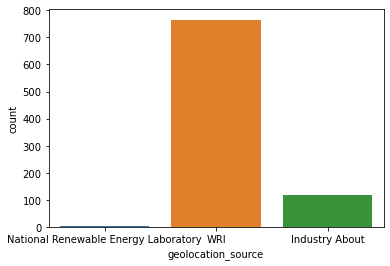
* 50 power plant use Biomass as its primary fuel
* 258 power plant use Coal as its primary fuel
* 69 power plant use Gas as its primary fuel
* 251 power plant use Hydro as its primary fuel
* 9 power plant use Nuclear as its primary fuel
* 20 power plant use Oil as its primary fuel
* 127 power plant use Solar as its primary fuel
* 123 power plant use Wind as its primary fuel

**other\_fuel1**

****

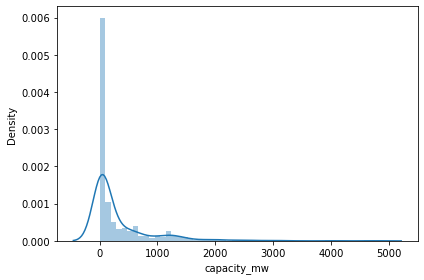
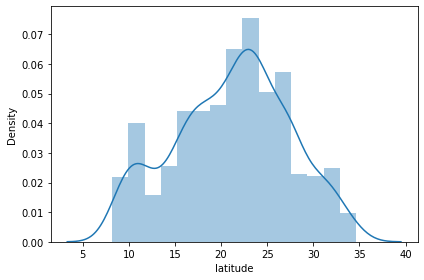
* 1 power plant use Cogeneration as its other fuel
* 125 power plant use Gas as its other fuel
* 781 power plant use Oil as its other fuel

**geolocation\_source**

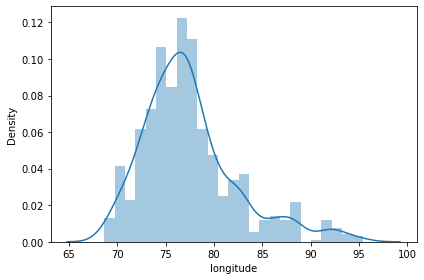
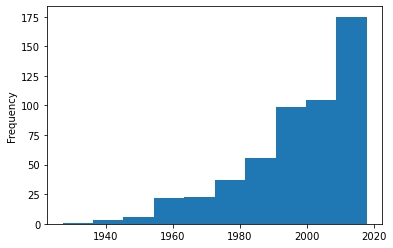
****

* For 124 power plants their geolocation source is Industry About
* For 9 power plants their geolocation source is National Renewable Energy Laboratory
* For 774 power plants their geolocation source is WRI

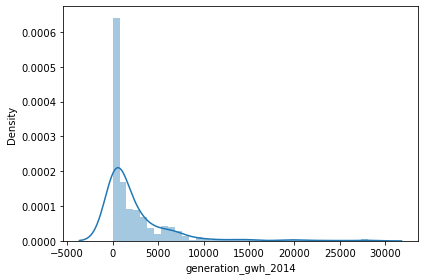
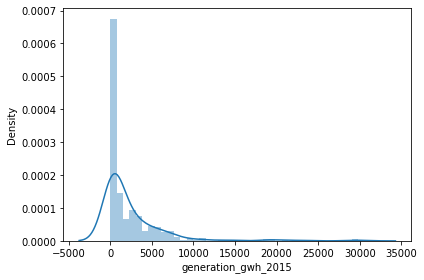
**capacity\_mw Latitude**

** **

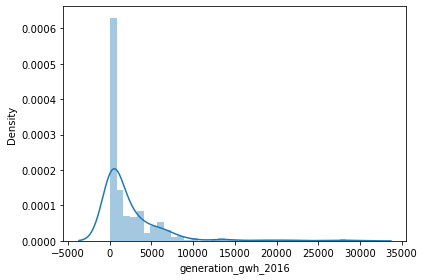
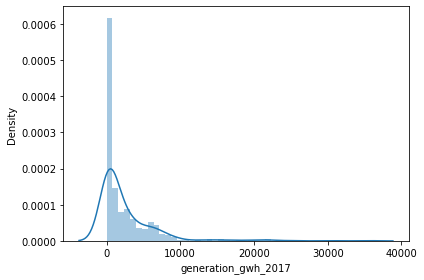
**Longitude commissioning\_year**

** **

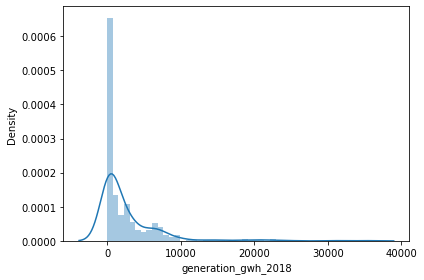
**generation\_gwh\_2014 generation\_gwh\_2015**

** **

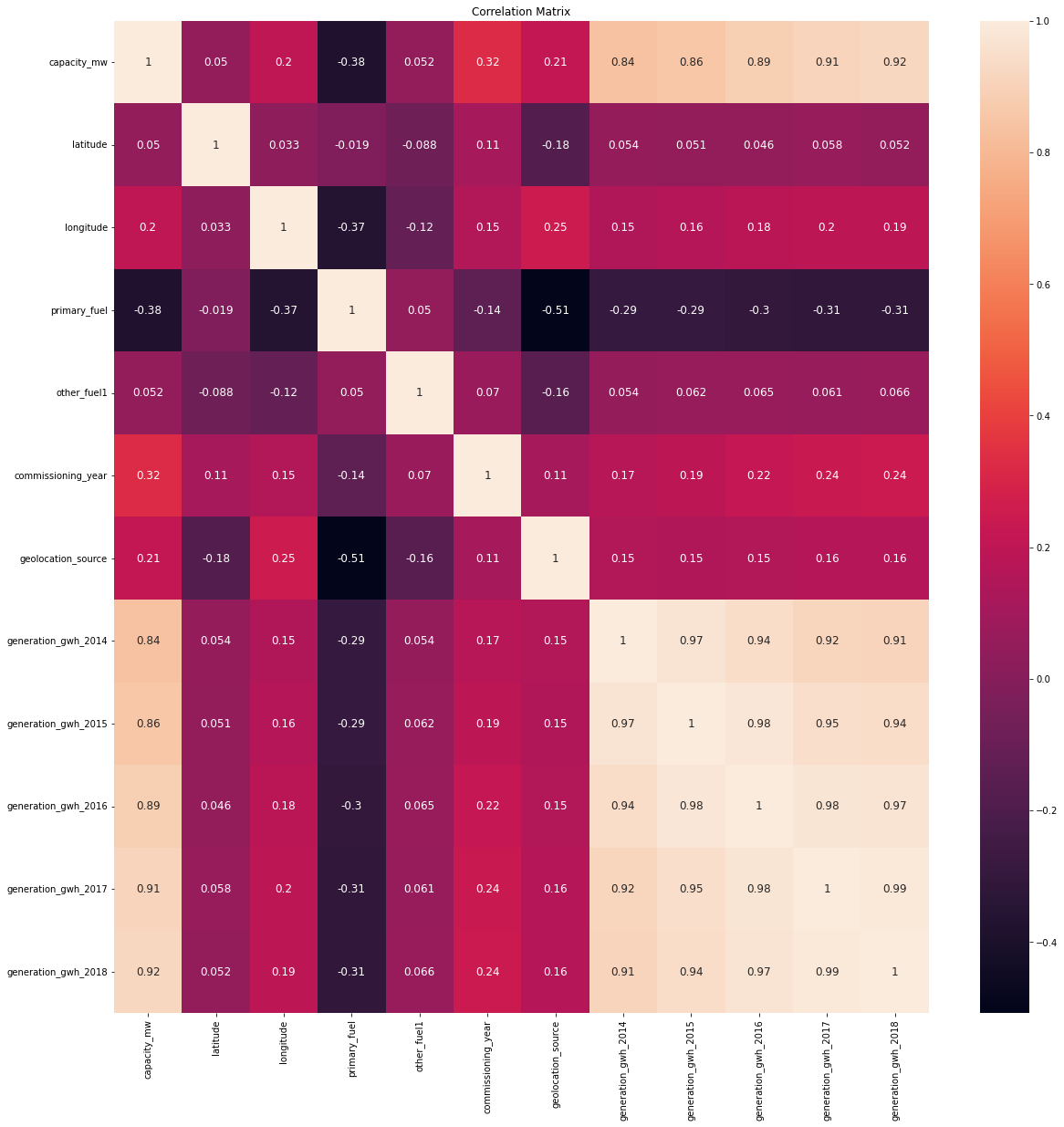
**generation\_gwh\_2016 generation\_gwh\_2017**

** **

**generation\_gwh\_2018**

****

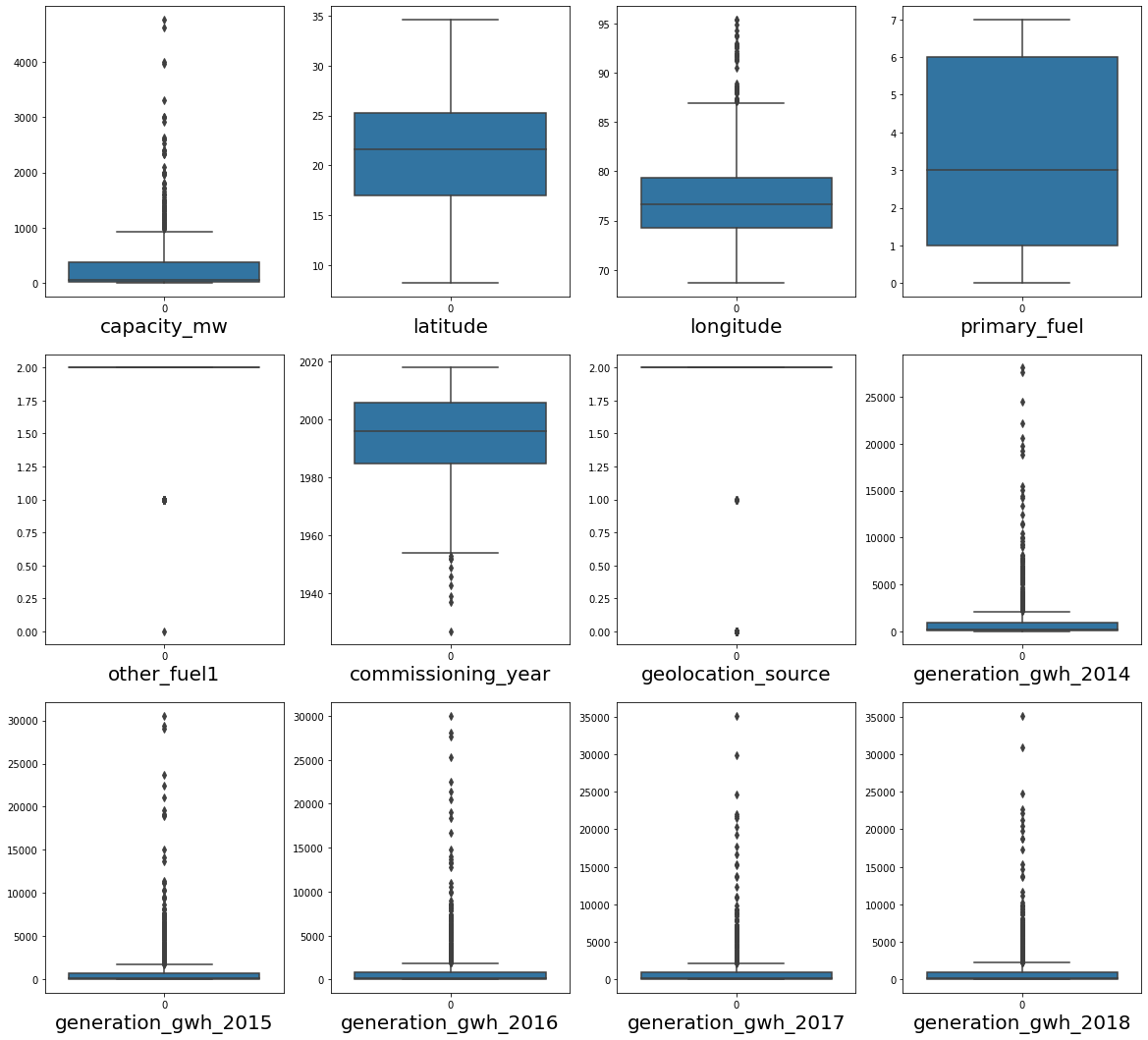
* **Correlation amongst the features**

****

* **Correlation between features and label-**

|  |  |
| --- | --- |
| **Label- primary\_fuel** | |
| **Feature** | **Correlation Coefficient** |
| geolocation\_source | -0.508032 |
| capacity\_mw | -0.380281 |
| longitude | -0.366274 |
| generation\_gwh\_2018 | -0.312769 |
| generation\_gwh\_2017 | -0.311821 |
| generation\_gwh\_2016 | -0.302689 |
| generation\_gwh\_2014 | -0.292698 |
| generation\_gwh\_2015 | -0.291749 |
| commissioning\_year | -0.142410 |
| latitude | -0.019233 |
| other\_fuel1 | 0.049981 |
| **Label- capacity\_mw** | |
| **Feature** | **Correlation Coefficient** |
| primary\_fuel | -0.380281 |
| latitude | 0.049789 |
| other\_fuel1 | 0.051538 |
| longitude | 0.204324 |
| geolocation\_source | 0.211651 |
| commissioning\_year | 0.324521 |
| generation\_gwh\_2014 | 0.839031 |
| generation\_gwh\_2015 | 0.857873 |
| generation\_gwh\_2016 | 0.887205 |
| generation\_gwh\_2017 | 0.906152 |
| generation\_gwh\_2018 | 0.918602 |

* **Checking for outliers-**

****

* The features with continuous data containing outliers-
* capacity\_mw
* longitude
* commissioning\_year
* generation\_gwh\_2014
* generation\_gwh\_2015
* generation\_gwh\_2016
* generation\_gwh\_2017
* generation\_gwh\_2018

**EDA Concluding Remark**

* The object data type of other fuel1, geolocation\_source and primary\_fuel needs to be encoded into numeric type
* The statistical analysis described that capacity\_mw, generation\_gwh\_2014, generation\_gwh\_2015, generation\_gwh\_2016, generation\_gwh\_2017, generation\_gwh\_2018 have their minimum values as 0, but that is not possible, hence they are being replaced by NaN
* The unknown values in other\_fuel1 and geolocation\_source are also replaced as NaN
* The graphical analysis suggests that the data in capacity\_mw, latitude, longitude, generation\_gwh\_2014, generation\_gwh\_2015, generation\_gwh\_2016, generation\_gwh\_2017 and generation\_gwh\_2018 is skewed.
* The correlation coefficient between the features and label primary fuel suggests the following-
  + geolocation\_source, capacity\_mw, longitude, generation\_gwh\_2018, generation\_gwh\_2017, generation\_gwh\_2016, generation\_gwh\_2014, generation\_gwh\_2015, commissioning\_year and latitude are negatively correlated to the label primary\_fuel
  + other\_fuel1 are 4% positively correlated to the label primary\_fuel showing a moderately weak bond
  + geolocation\_source is 50.8% negatively correlated with primary\_fuel showing a strong bond
  + latitude is 0.1% negatively correlated with primary\_fuel showing an extremely weak bond
  + The correlation coefficient between the features and label capacity\_mw suggests the following-
  + geolocation\_source, other\_fuel1, primary\_fuel, longitude, generation\_gwh\_2018, generation\_gwh\_2017, generation\_gwh\_2016, generation\_gwh\_2014, generation\_gwh\_2015, commissioning\_year and latitude are positively correlated to the label capacity\_mw
  + primary\_fuel are 38% negatively correlated to the label capacity\_mw showing a moderately strong bond
  + generation\_gwh\_2018 is 91.86% positively correlated with capacity\_mw showing an extremely strong bond
  + latitude is 4.98% positively correlated with capacity\_mw showing an moderately weak bond

**Preprocessing Pipeline**

* The object data is encoded using the label encoder, . whereby the string encoded in alphabetical manner as follows-
* primary\_fuel-
* 0- Biomass
* 1- Coal
* 2- Gas
* 3- Hydro
* 4- Nuclear
* 5- Oil
* 6- Solar
* 7- Wind
* other\_fuel1-
* 0- Cogeneration
* 1- Gas
* 2- Oil
* 3- Null
* geolocation\_source-
* 0- Industry About
* 1- National Renewable Energy Laboratory
* 2- WRI
* 3- Null
* The null values are imputed into meaningful data using the KNN Imputer
* The outliers are removed using the zscore method. The data loss is 5.18%, which is tolerable, The size of the reformed dataset is 860 rows and 12 columns
* The data set is divided into features (x) and label (y)
* The skewness observed in graphical analysis, is confirmed as follows-

|  |  |
| --- | --- |
| **Feature Name** | **Skewness Value** |
| generation\_gwh\_2014 | 2.749187 |
| generation\_gwh\_2015 | 2.616752 |
| generation\_gwh\_2016 | 2.571109 |
| generation\_gwh\_2018 | 2.536255 |
| generation\_gwh\_2017 | 2.477173 |
| capacity\_mw | 2.010388 |
| longitude | 0.974769 |
| latitude | -0.103976 |
| commissioning\_year | -0.498210 |
| geolocation\_source | -1.982817 |
| other\_fuel1 | -2.112517 |

* The skewness is removed using the power transformer
* The data was then scaled using the standard scaler
* The best random state is traced for the train test split.
* This random state is applied to divide the features in train and test sets

**Building Machine Learning Models**

**Prediction 1- Primary-fuel**

**As the data of primary fuel is categorical, classification models are to be used.**

The following machine learning models are used-

* **Logistic Regression-**

Accuracy 75.0

[[ 5 0 0 2 0 0 2]

[ 1 46 0 7 0 0 0]

[ 0 3 1 4 0 0 0]

[ 1 6 0 29 0 0 6]

[ 1 0 0 0 0 0 4]

[ 0 0 0 0 0 27 0]

[ 2 1 0 3 0 0 21]]

precision recall f1-score support

0 0.50 0.56 0.53 9

1 0.82 0.85 0.84 54

2 1.00 0.12 0.22 8

3 0.64 0.69 0.67 42

5 0.00 0.00 0.00 5

6 1.00 1.00 1.00 27

7 0.64 0.78 0.70 27

accuracy 0.75 172

macro avg 0.66 0.57 0.56 172

weighted avg 0.74 0.75 0.73 172

* **Decision Tree Classifier-**

Accuracy 79.65116279069767

[[ 8 0 0 0 0 1 0 0]

[ 2 38 5 5 1 0 0 3]

[ 0 0 2 4 0 0 0 2]

[ 1 1 0 37 0 1 0 2]

[ 0 0 0 0 0 0 0 0]

[ 0 0 1 1 0 2 0 1]

[ 0 0 0 0 0 0 26 1]

[ 1 0 0 2 0 0 0 24]]

precision recall f1-score support

0 0.67 0.89 0.76 9

1 0.97 0.70 0.82 54

2 0.25 0.25 0.25 8

3 0.76 0.88 0.81 42

4 0.00 0.00 0.00 0

5 0.50 0.40 0.44 5

6 1.00 0.96 0.98 27

7 0.73 0.89 0.80 27

accuracy 0.80 172

macro avg 0.61 0.62 0.61 172

weighted avg 0.82 0.80 0.80 17

# Random Forest Classifier-

Accuracy 84.30232558139535

[[ 7 1 0 0 0 0 1]

[ 2 46 0 4 0 0 2]

[ 0 3 3 2 0 0 0]

[ 0 2 0 38 0 0 2]

[ 0 0 0 2 2 0 1]

[ 0 0 0 0 0 27 0]

[ 2 1 0 2 0 0 22]]

precision recall f1-score support

0 0.64 0.78 0.70 9

1 0.87 0.85 0.86 54

2 1.00 0.38 0.55 8

3 0.79 0.90 0.84 42

5 1.00 0.40 0.57 5

6 1.00 1.00 1.00 27

7 0.79 0.81 0.80 27

accuracy 0.84 172

macro avg 0.87 0.73 0.76 172

weighted avg 0.85 0.84 0.84 172

# SVC-

Accuracy 76.74418604651163

[[ 6 0 0 1 0 0 2]

[ 1 48 0 5 0 0 0]

[ 0 1 0 6 0 0 1]

[ 1 6 0 30 0 0 5]

[ 0 0 1 0 0 0 4]

[ 0 0 0 0 0 27 0]

[ 5 1 0 0 0 0 21]]

precision recall f1-score support

0 0.46 0.67 0.55 9

1 0.86 0.89 0.87 54

2 0.00 0.00 0.00 8

3 0.71 0.71 0.71 42

5 0.00 0.00 0.00 5

6 1.00 1.00 1.00 27

7 0.64 0.78 0.70 27

accuracy 0.77 172

macro avg 0.52 0.58 0.55 172

weighted avg 0.72 0.77 0.74 172

**Cross Validation of each model-**

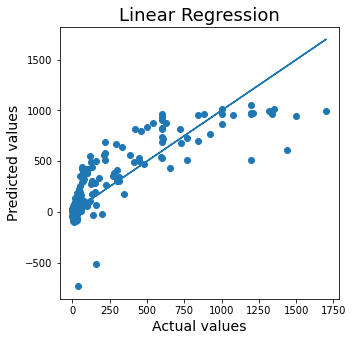
* Cross validation score of Logistic Regression: 0.44302325581395346
* Cross validation score of Decision Tree Classifier: 0.6953488372093023
* Cross validation score of Random Forest Classifier: 0.776744186046511
* Cross validation score of SVC model: 0.3848837209302326

**Accuracy and cross validation both suggest that random forest classifier works the best**

**Hyperparameter tuning of Random Forest Classifier with GrideSearchCV with criterion as entropy, max\_depth as 8 and max\_features as sqrt, gives final accuracy as 82%**

# Prediction 2- capacity\_mw

**As the data of capacity\_mw is continuous, regression models are to be used.**



The Best Fit Line shows the least values of residuals and covering most of the data point, showing a good fit for our model

The following machine learning models are used-

# Linear Regression with Lasso Regularization using GridSearchCV, using alpha as 1 and random state as 0-

# R2 score- 66.8

# Cross validation score- 60.99

* **Random Forest Regressor hypertuned with GridSearchCV, taking criterion as mae and max\_features as sqrt**

# R2 score- 84.79

# Cross validation score- 85.08

**Concluding Remarks**

* Reducing the number of columns was a good decision as it helped in better and faster processing
* The data needs to be analysed well, and all outliers, skewness as well as data misinformation must be rectified before fitting it into the models
* The best random state is preferred as it enhances the model efficiency
* The models must assessed with cross validation to check for overfitting issues
* Regularization and hyperparameter tuning is very helpful to improve the accuracy of the model
* **For prediction 1 (primary fuel) the hyperparameter tuned random forest classifier is the best working model, providing an accuracy of 82%**
* **For prediction 2 (capacity\_mw) the hyperparameter tuned random forest regressor is the best working model, providing a R2 score of 84.79% and cross validation score of 85.08%.**