DATA SCIENCE INTERNSHIP ASSIGNEMENT

(MECHADEMY)

1. DATA LOADING AND EDA

Imported Necessary libraries for development and loaded the data using CSV() by mentioning the file path and verified given below,

a. Null values
b. Data Type
c. Total Features
d. Outliers

2. FEATURE ENGINEERING AND FEATURE SELECTION

- As a part of feature engineering, I have found there are some NULL VALUES and replaced them by using **mean().**
- And also converted the object type data features into float type **.to_numeric().**
- In feature selection process, i came to know that specially mentioned values random_variable1 and random_variable2 are not mandatory to use and built a model.
- I concluded not to use them by using Random Forest Regressor, correlation and select k best methods, by analysing all the above methods and statistically i have decided to not to use in model building as shown in given below data and pictures,

CORRELATION MATIRX :

	random_variable1	random_variable2
random_variable1	1.000000	0.280647
random_variable2	0.280647	1.000000
Equipment energy consumption	-0.015383	-0.010770

	Equipment energy consumption
random_variable1	-0.015383
random_variable2	-0.010770
Equipment energy consumption	1.000000

SELECT K BEST :

Feature		Score
Lighting energy	-	49.371695
Outdoor humidity	-	23.335567
Outdoor temperature	-	16.944951
Random variable1	-	3.989646
Atmospheric pressure	-	2.974449
Random variable2	-	1.955420
Visibility index	-	0.000006

- From the above data, we can understand there is no high relation of random_variable1 and random_variable2 with other features in the data set.
- And based on correlation method and technically I have dropped features except lighting energy, outdoor temperature, atmospheric pressure, outdoor humidity, visibility index.

CORRELATION B/W TARGETED FETURE AND REMAINING FEATURES :

Correlation between timestamp and equipment energy consumption: -0.0039

Correlation between **lighting energy and equipment energy consumption: 0.0540**

Correlation between zone1_temperature and equipment energy consumption: 0.0174

Correlation between zone1_humidity and equipment energy consumption: 0.0253

Correlation between zone2_temperature and equipment energy consumption: 0.0398

Correlation between zone2_humidity and equipment energy consumption: -0.0037

Correlation between zone3_temperature and equipment energy consumption: 0.0362

Correlation between zone3_humidity and equipment energy consumption: 0.0063

Correlation between zone4_temperature and equipment energy consumption: 0.0163

Correlation between zone4_humidity and equipment energy consumption: -0.0031

Correlation between zone5_temperature and equipment energy consumption: 0.0085

Correlation between zone5_humidity and equipment energy consumption: 0.0076

Correlation between zone6_temperature and equipment energy consumption: 0.0304

Correlation between zone6_humidity and equipment energy consumption: -0.0184

Correlation between zone7_temperature and equipment energy consumption: 0.0069

Correlation between zone7_humidity and equipment energy consumption: -0.0065

Correlation between zone8_temperature and equipment energy consumption: 0.0189

Correlation between zone8_humidity and equipment energy consumption: -0.0229

Correlation between zone9_temperature and equipment energy consumption: 0.0038

Correlation between zone9_humidity and equipment energy consumption: -0.0217

Correlation between **outdoor temperature and equipment energy consumption: 0.0317**

Correlation between **atmospheric pressure and equipment energy consumption: -0.0133**

Correlation between **outdoor humidity and equipment energy consumption: -0.0372**

Correlation between wind speed and equipment energy consumption: 0.0110

Correlation between **visibility index and equipment energy consumption: 0.0000**

Correlation between dew point and equipment energy consumption: -0.0031

Correlation between random_variable1 and equipment energy consumption: -0.0154

Correlation between random_variable2 and equipment energy consumption: -0.0108

- Some data points are better correlated, but even though they are removed as they are specifically related to particular zone. So it doesn't impact much higher, as data differs from zone to zone, below given data features are selected to train with model.

a. Lighting energy
b. Outdoor temperature
c. Atmospheric pressure
d. Outdoor humidity
e. visibility_ index

3. Model building and Evaluation

Trained with two Alogorithms,

a. Linear Regression:
Mean Squared Error: 32093.201253415
R^2 Score: 0.0028384743411244973

b. Random Forest Regressor:
Mean Squared Error: 31770.854197057855
R^2 Score : 0.012854056145215731

- By comparing two models I have selected Random Forest Regressor, as it gave better result than Linear regression

4. My observations and Recommendations

- I observed, some features like Lighting energy, Out Door temperature, Out Door humidity, atmospheric pressure and visibility index is related to energy consumption.

- Even though features like zone temperature and humidity is correlated with energy Consumption, I decided not to use to train model, as we know things like temperature, humidity, visibility index differs from one zone to another.

- We should consume energy based on features like Lighting energy, Out Door temperature, Out Door Humidity, atmospheric pressure and visibility index, as they impact energy consumption.