PREDICTION USING MULTIPLE LINEAR MODEL EXTENSIONS FOR STEEL ENERGY CONSUMPTION DATASET

By:

- **Yashwantej Dyavari Shetty.**

Content

- Objective and Dataset Description
- Data visualization
- Linear regression and implementation
- Generalized Additive Models
- Random Forest
- Conclusion

Main Objective

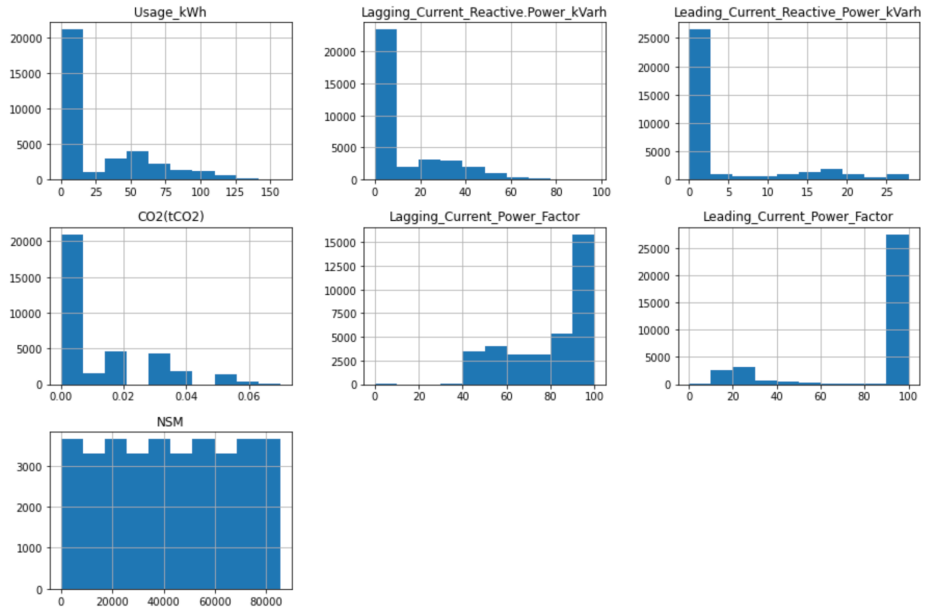
- The fast development of urban advancement in the past decade requires reasonable and realistic solutions for transport, building infrastructure, natural conditions, and personal satisfaction in smart cities.
- This project presents and explores predictive energy consumption Usage_kWh as the target variable using models for a smart small-scale steel industry.
- Energy consumption data is collected using IoT-based systems and used for prediction.
- Data used include the lagging and leading current reactive power, the lagging and leading current power factor, carbon dioxide emissions, and load types.
- Five statistical algorithms are used for energy consumption prediction:(a) Multiple linear regression, (b) Generalized Additive Models, (c) Random Forest and Boosting.
- Root mean squared errors are used to measure the prediction efficiency of the models.

Dataset

- The dataset used is the Steel_industry_data dataset
- The features of the dataset are as follows :
- ‘date', 'Usage_kWh', 'Lagging_Current_Reactive.Power_kVarh', 'Leading_Current_Reactive_Power_kVarh', 'CO2(tCO2)’, 'Lagging_Current_Power_Factor', 'Leading_Current_Power_Factor', 'NSM’, 'WeekStatus', 'Day_of_week', 'Load_Type’.
- Number of Observations: 35040
- Number of Predictors: 10
- The target variable is 'Usage_kWh’ and the remaining features are independent features.
- Usage_kWh: This refers to the total amount of energy used in kilowatt-hours (kWh) during a given period.
- WeekStatus: This indicates whether the given day is a weekday or a weekend.

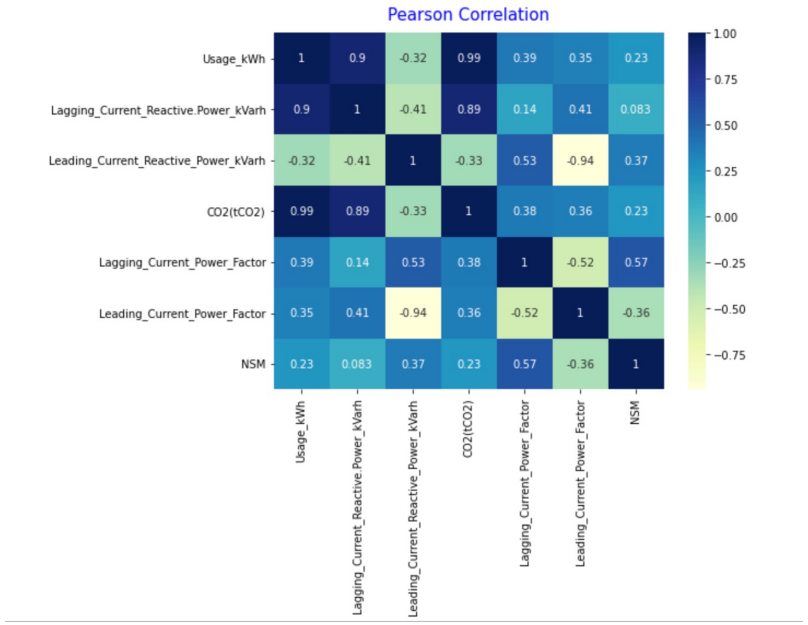
Dataset

- Lagging_Current_Reactive.Power_kVarh: This refers to the amount of reactive power in kilovolt-ampere reactive hours (kVarh) that is lagging behind the voltage waveform in an AC electrical system.
- Leading_Current_Reactive_Power_kVarh: This refers to the amount of reactive power in kilovolt-ampere reactive hours (kVarh) that is leading the voltage waveform in an AC electrical system.
- CO2(tCO2): This refers to the amount of carbon dioxide (CO2) emissions in metric tons (tCO2) associated with the energy consumption during the given period.
- Lagging_Current_Power_Factor: This refers to the ratio of real power to apparent power in an AC electrical system when the reactive power is lagging behind the voltage waveform.
- Leading_Current_Power_Factor: This refers to the ratio of real power to apparent power in an AC electrical system when the reactive power is leading the voltage waveform.
- NSM: This refers to the number of seconds since midnight.
- Day_of_week: This refers to the day of the week (e.g., Monday, Tuesday, etc.) corresponding to the given period.
- Load_Type: This refers to the type of energy load, such as residential, commercial, or industrial, associated with the given energy consumption data



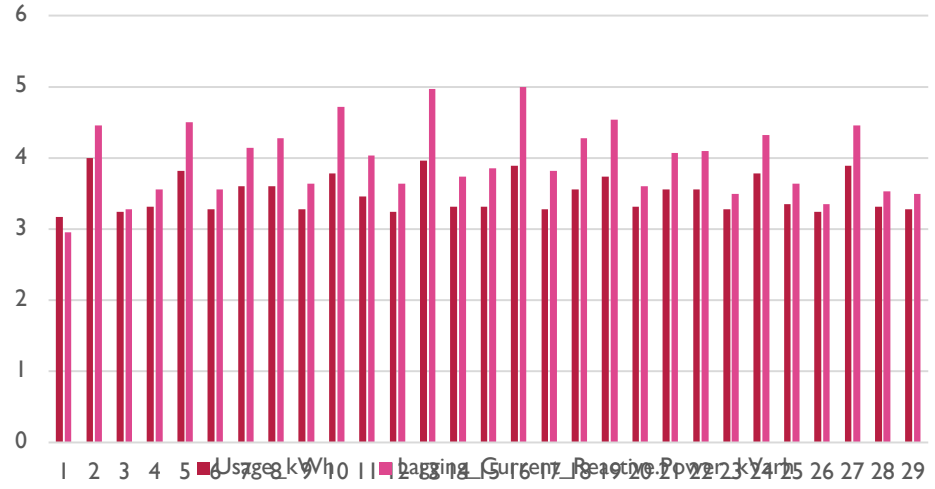
Data Pre-Processing & Data Understanding

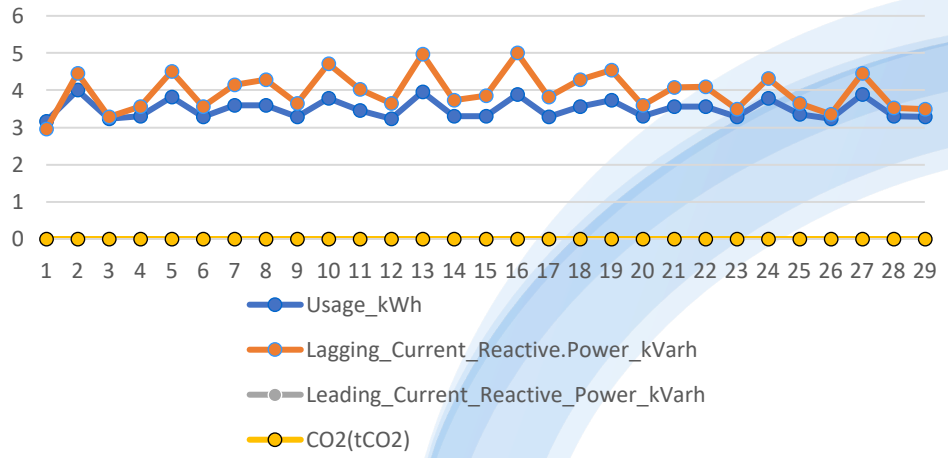
- **Steel_industry_data dataset:-**
- For data Preprocessing purposes, the existence of NULL values is analyzed, and observed that no null values in the dataset.
- **Correlation among features of the dataset :**
- Correlation explains how one or more variables are related to each other. These variables can be input data features that have been used to forecast our target variable.
- Correlation is a statistical technique that determines how one variable moves/changes in relation to the other variable. It gives us an idea about the degree of the relationship between the two variables. It's a bi-variate analysis measure that describes the association between different variables.

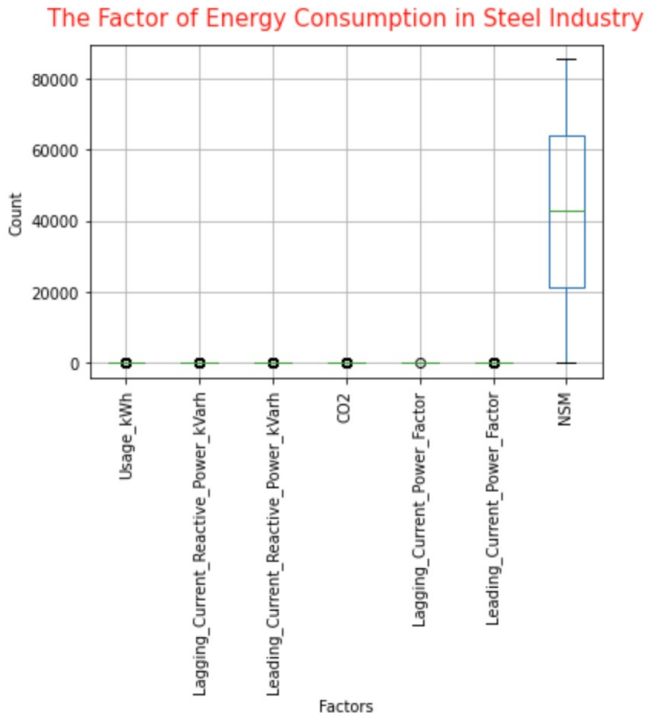


Visualization of the Data

- Data visualization for Usage_kWh vs Lagging_Current_Reactive.Power_kVarh for a sample of data from dataset and Usage_kWh vs Lagging CR power, Leading CR power, CO2







- Usage_kWh data is varying with the different combinations of remaining independent features
- data = read.csv(“Steel_industry_data.csv")
- head(data)

RELATIONSHIP BETWEEN THE VARIABLES OF DATASET

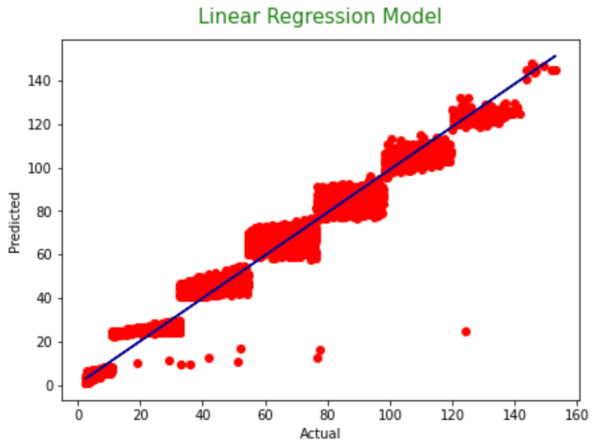
LINEAR REGRESSION IMPLEMENTATION DETAILS

- Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.
- Linear regression is of the form, Y = a + bx, where a is an intercept, b is the coefficient and x is the independent feature
- The task is to reduce the difference between the actual value and the predicted value using the ML model.
- Minimize (Ypred, Yactual), by taking optimal parameters

R CODE TO IMPLEMENT MULTIPLE LINEAR REGRESSION MODEL

```r
library(caret)

# Read the CSV file into a data frame

df = read.csv("/Users/yashds/Documents/AdvStat2/Project/Steel_industry_data.csv")
data("df")
head(df)
# Split the data into training and test sets
X <- subset(df, select = -Usage_kWh)
y <- df$Usage_kWh
set.seed(42)
split <- createDataPartition(y, p = 0.7, list = FALSE)
X_train <- X[split, ]
X_test <- X[-split, ]
y_train <- y[split]
y_test <- y[-split]

# Fit a linear regression model on the training set
model <- lm(y_train ~ ., data = X_train)

# Make predictions on the test set
y_pred <- predict(model, newdata = X_test)

# Evaluate the model
rmse <- sqrt(mean((y_pred - y_test)^2))
cat("RMSE:", rmse, "\n")
```

Multiple Linear Regression	RMSE
Usage_kWh	4.21537



Linear Regression is a very suitable model to predict energy consumption from the Steel Industry because has a good accuracy score of 98% with Root Mean Squared Error:4.215375315598361

MULTIPLE LINEAR REGRESSION

- Models the linear relationship between a single dependent continuous variable and more than one independent variable.
- Multiple linear regression is a subset of polynomial regression. An nth degree polynomial in x is used to model the relationship between the independent variable x and the dependent variable y.
- Non-linear data cannot be fitted using linear regression (underfitting). As a result, we enhance the model's complexity and employ Polynomial regression, which better fits the data. (in the form y = a0 + a1x1 + a2x12 +... + anx1n), where a0, a1,….are parameters and x1, x2,…. Xn are features
- With an accuracy of approximately 98%, linear regression is performing better in prediction and also with rmse value of 4.21

Generalized Additive Models

Generalized additive Models	RMSE
Usage_kWh	0.0505

- The results of a linear regression model are the weighted sum of variables. This is a model weakness, but it is also a strength. However, when modelling with data that does not have a Gaussian distribution, the results of a simple linear model can be nonlinear. To improve the model, we can make a number of changes.
- GAM(Generalized Additive Model) is an extension of linear models. As we know, the formula of linear regression is:
- $y=\beta 0+\beta 1X1+\ldots+\beta pxp+\epsilon$

Random Forest

- Random Forest is a well-known and powerful machine-learning algorithm. Bootstrap Aggregation, also known as bagging, is a type of ensemble machine-learning algorithm.
- The bootstrap method for estimating statistical quantities from samples.
- The Bootstrap Aggregation algorithm for creating multiple different models from a single training dataset.
- The Random Forest algorithm makes a small tweak to Bagging and results in a very powerful classifier.

```r
```{r}
library(caret)

# Split the data into training and test sets
X <- subset(df, select = -Usage_kWh)
y <- df$Usage_kWh
set.seed(42)
split <- createDataPartition(y, p = 0.7, list = FALSE)
X_train <- X[split, ]
X_test <- X[-split, ]
y_train <- y[split]
y_test <- y[-split]

# Fit a random forest model on the training set
library(randomForest)
model <- randomForest(y_train ~ ., data = X_train)

# Make predictions on the test set
y_pred <- predict(model, newdata = X_test)

# Evaluate the model
rmse <- sqrt(mean((y_pred - y_test)^2))
cat("RMSE:", rmse, "\n")
```
```

```
randomForest 4.7-1.1
Type rfNews() to see new features/changes/bug fixes.

Attaching package: 'randomForest'

The following object is masked from 'package:ggplot2':

    margin

RMSE: 1.941737
```

Random Forest	RMSE
Usage_kWh	1.9417

Boosting

- **Types of Boosting:-**
- AdaBoost , Gradient Boosting, XGBoosting
- Boosting is an ensemble method that builds on weak learners one by one until one last strong learner emerges.
- A weak learner is a model that isn't extremely accurate or doesn't take into consideration numerous predictions.
- Boosting can effectively convert weak learners into strong learners by generating a weak model, drawing conclusions about the various feature importances and parameters, and then using those conclusions to build a new, stronger model. BOTH classification and regression issues can benefit from boosting.

```r
```{r}

library(caret)

# Read the CSV file into a data frame
df <- read.csv("Steel_industry_data.csv", header = TRUE)

# Split the data into training and test sets
X <- subset(df, select = -Usage_kWh)
y <- df$Usage_kWh
set.seed(42)
split <- createDataPartition(y, p = 0.7, list = FALSE)
X_train <- X[split, ]
X_test <- X[-split, ]
y_train <- y[split]
y_test <- y[-split]

# Fit a boosting model with gradient boosting machines (gbm)
library(gbm)
model_boost <- gbm(y_train ~ ., data = X_train, n.trees = 100, shrinkage = 0.01, interaction.depth = 3)

# Make predictions on the test set
y_pred_boost <- predict(model_boost, newdata = X_test, n.trees = 100)

# Evaluate the model
rmse_boost <- sqrt(mean((y_pred_boost - y_test)^2))
cat("RMSE (boosting):", rmse_boost, "\n")
```
```

Boosting	RMSE
Usage_kWh	1.073

Conclusion

- After training all models for the Steel industry dataset. We have evaluated with RMSE. Based on that we have given the report.
- The results show that for Usage_kWh Multiple Linear Regression model performed well provides best results with lower error values and this model can be used for the development of energy efficient structural design which helps to optimize the energy consumption and policy making in smart cities.