**Energy Star Score Prediction - Regression**

**Goal** : Predict the Energy Star score of a building and find the factor influencing the score

**Quick Data Review**

1. The dataset has number of rows 11746 and columns 60.
2. Number of obj 48 ,number of integers 6 and number of floats are 6.
3. Most of the fetures are objects with 'non available'. Need to check does the features is made as object because of non available though it has numbers in it.

**Data cleaning and formatting**

1. Replace all Not Available into NaN ( np.nan)
2. convert the datatypes of features with [ 'ft²', 'therms', 'kBtu', 'Metric Tons CO2e' , 'gal' , 'kWh' , 'Score' from object to float.
3. Find the Missing data and Remove columns with more than 50% of data missing

**Exploratory data analysis**

1. Target variable Distribution - ENERGY STAR Score is not uniformly distributed
2. The Site EUI (kBtu/ft²)' has outlier with value of 869265 .
3. So outlier 3.0 times the IQR below the first quartile or above the third quartile were removed.
4. "Largest Property Use Type"has some significant effect on "ENERGY STAR Score”.
5. Find Correlations between Features and Target and plot the Pairplot

**Feature engineering and selection**

1. Log transformation of the numerical variables.
2. One-hot encode the categorical variables
3. Remove collinear features with threshold of 0.6 – multicollinearity
4. Replace the inf and -inf with nan (required for later imputation)
5. Split Into Training and Testing Sets

**Establish a Baseline**

1. Performance metric - MAE
2. Baseline guess - median training value
3. Baseline Performance – np.mean(abs(y\_test, baseline\_guess))

#### **Data Pre-processing**

##### Imputing the Missing values by median

##### Make sure all values are finite

##### Scaling Features – MinMaxScaler – 0-1

**Compare several machine learning models on a performance metric**

1. Develop Programs of application of ML models ( MAE & Fit & Evaluate)
2. Apply the Linear Regression, SVR, Random Forest Regression, Gradiant Boosting.K-Nearest Neighbors Regression
3. Compare all the ML models to the above program and find the best
4. Gradiant Boosting showed the best MAE score of 10.011.

**Perform hyperparameter tuning on the best model to optimize it for the problem**

1. Random search with CV is a good method to narrow down the possible hyperparameters
2. Grid search with CV is applied on random search results
3. grid\_search.cv\_results\_ into a results dataframe
4. Plot the training and testing error vs number of trees

plot results['param\_n\_estimators'], -1 \* results['mean\_train\_score'] – Training Error

plot results['param\_n\_estimators'], -1 \* results['mean\_test\_score'], – Testing Error

1. Conclude from above graph Overfitting/ Underfitting/good fit

**Evaluate the best model on the testing set**

1. Fit the grid\_search.best\_estimator\_as **final model** on train data
2. Predict on the test data
3. The final model of Gradient Boosting with hyperparameter Tuning has performed well than default GB. And the final model MAE is 9.07
4. Check distribution of residual of final model – It should be  normally distributed

**Interpret the model results to the extent possible**

1. Site EUI & Wheather Normalized Site Electricity Intensity has got more importance in prediction and others have very less role
2. final\_model.feature\_importances\_ provides list of important features
3. Select the top 10 features and fit the **final model** on that 10 features and compare whether the MAE reduced futher or not. In our case

**Draw conclusions and write a well-documented report**

1. With the provided data, machine learning model can predict the Energy Star Score of a building to within 10 points.
2. Electricity Use Intensity is an important paramater for determining the Energy Star Score