Project Description: Analyzing Building Data and Building a Predictive Energy Efficiency Model

1. Introduction

This project focuses on analyzing building data related to energy and water consumption and building a predictive model to estimate each building's ENERGY STAR Score. The project is based on real data from New York City buildings and applies a series of practical steps to process and explore the data, train a robust model, and interpret its results.

2. Data Processing

Data Loading:

Reading a CSV file containing energy and water consumption information and building data.

Data Cleaning:

Converting text values ​​"Not Available" to missing values ​​(NaN).

Managing Numeric Columns:

Converting numeric text to numbers.

Filling missing values ​​with the median.

Handling outliers using the IQR (Interquartile Range) method.

Handling Time Columns: Converting them to the date data type and filling missing values ​​with the median.

Handling Geographic Columns: Deleting rows missing latitude and longitude coordinates.

Handling Text Columns: Fill Missing Values ​​with "Unknown."

Explore Data:

Graph the ENERGY STAR Score distribution to understand how scores are spread across buildings.

3. Prepare Data for the Model

Numeric Features:

Take the logarithm of some columns to reduce the impact of outliers.

Categorical Features:

Convert categorical columns such as "Borough" and "Largest Property Use Type" to a numerical representation using One-Hot Encoding.

Target Column:

The target column is the ENERGY STAR Score.

Remove the target column from the feature set.

Data Splitting:

Split the data into a training set (70%) and a test set (30%).

Fill Missing Values ​​and Normalize Data:

Use SimpleImputer to fill missing values ​​in features.

Use MinMaxScaler to normalize data between 0 and 1.

4. Model Building and Training

Model used: GradientBoostingRegressor, a powerful model for estimating numerical values.

Main Settings:

loss='absolute\_error'

max\_depth=5

n\_estimators=500

Other parameters to adjust the depth and number of samples per split.

Training: The model learns from the training data after processing and normalization.

5. Model Evaluation and Feature Importance

Extract the importance of features to determine the factors that most influence the prediction.

Graphically display the top 10 features to illustrate the impact of each feature.

Analysis of the sample with the largest prediction error to identify model weaknesses.

6. Visually Interpret the Model

Individual Tree View:

Extract a single tree from the Gradient Boosting model and convert it to a .png image to understand internal decisions.

Using LIME:

Interpret individual predictions, especially the sample with the largest error.

Visually display the impact of each feature on the prediction.

7. Conclusion

The project covers the entire process from data processing → data exploration → model training → interpretation of results.

The model:

It is robust and realistic, based on real data.

It provides practical information on the factors most affecting the energy efficiency of buildings.

Practical Uses:

It helps engineers and urban planners evaluate building efficiency.

It provides results that can be explained to end users or decision-makers using interpretation tools such as LIME.

Note: The project can be further developed using additional techniques such as:

Hyperparameter tuning (HPT).

Using cross-validation to evaluate the model more accurately.

Comparing the model with other models such as Random Forest or XGBoost.