Energy Consumption Prediction

By- Anushika Mishra(22110029)

Abstract

Climate change has become an issue of prominence owing to the increasing damage being caused to the environment. To address this problem we can work on mitigating the harmful emissions of greenhouse gasses and on adapting to the unavoidable consequences of climate change. To work on the former, we need to make changes in the electrical systems, buildings, land use, industry and transportation. Since buildings contribute a large percentage of energy and process related emissions, our work will be to look at buildings and predicting its Site Energy Usage Intensity (Site EUI) values based on the building details and weather data at the building's location so that we can know how making some easy fixes or state-of-the-art solutions can help change building energy consumption. In this work, we first analyze the data according to their facility type, make inferences, and try different approaches and machine learning techniques for making building energy consumption predictions. To implement this, we used different approaches to pre-processing the data and used two models: Dense Neural Network, and CatBoost Regressor. For baseline, we implemented linear regression and random forests, and the CatBoost Regressor is able to beat both the baseline scores.

Acknowledgements: I want to thank Prof. Anirban Dasgupta, Computer Science and Engineering, IIT Gandhinagar and the TAs in this course for their help and support throughout the project.

Introduction

As we are well aware, climate change is a global, pressing and multifaceted issue that is heavily influenced by infrastructure and energy policy. To address the issue of climate change, we must work in two directions, i.e, mitigating the emissions of greenhouse gas and adapting to unavoidable climate change consequences. The former of these methods requires a change in the electricity systems, buildings, land use, industry, and transportation. The International Energy Agency (IEA) reports the lifecycle of buildings from construction to demolition contributed for 37% of global energy-related and process-related emissions. However, we can drastically reduce this energy consumption by some easy-to-implement solutions and some state-of-the-art strategies. Example of such solutions is retrofitting some buildings could reduce heating and cooling requirements by 50 - 90 %. We can achieve overall cost savings and other benefits by employing energy efficient measures. To look at these energy efficient measures, we must first understand and analyze how the energy efficiency of buildings is related to the details of the building and climate data. This is where Energy Consumption Prediction comes into the picture.

The entire code for this project can be found in this repository:
https://github.com/AnushikaMishra/CS328

- **Problem Statement:** Given the data of a building and climate features, predict the Site Energy Usage Intensity (EUI) of the building.

Methodology

1. Dataset :

The dataset is taken from Kaggle from the WiDS Datathon. This dataset was created in collaboration with Climate Change AI (CCAI) and Lawrence Berkeley National Laboratory (Berkeley Lab). The prediction task in hand roughly involves one hundred thousand observations of building energy usage records that have been collected over a span of seven years across a number of states within the United States. The dataset consists of building characteristics like:

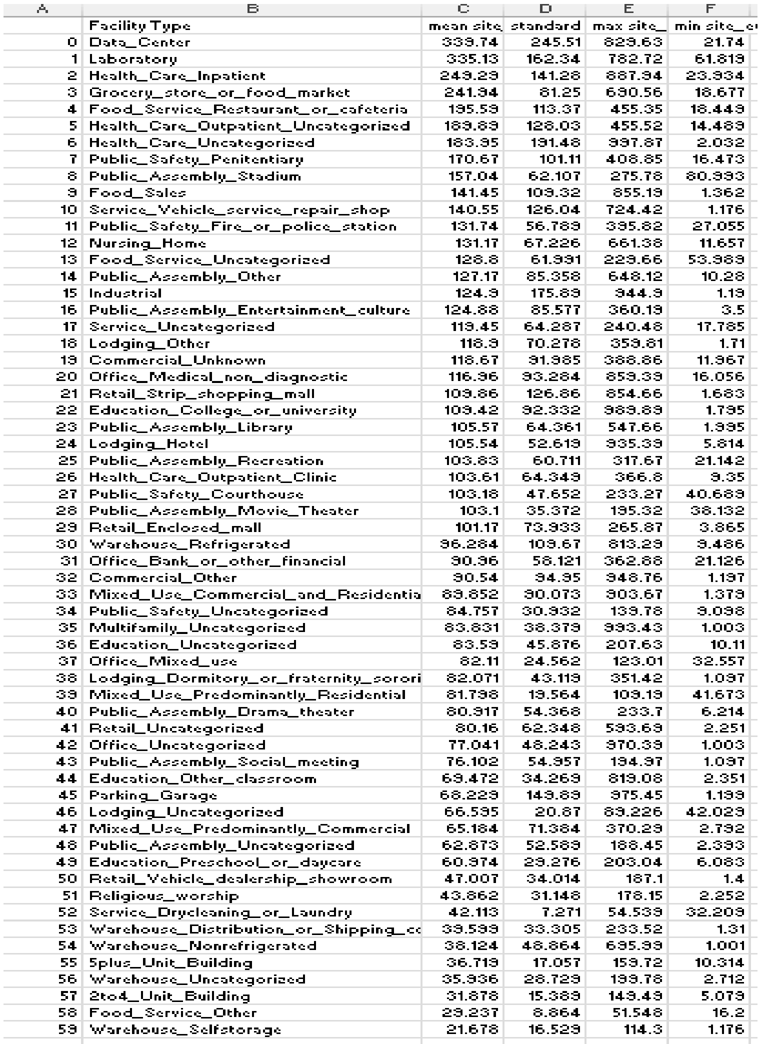
- building class
- facility type
- floor area
- year built
- energy star rating
- elevation of the building location

It also contains weather data like:

- heating and cooling degree days
- precipitation in inches
- snowfall in inches
- snow depth in inches
- average temperature
- counts of days below and above a certain range of temperatures
- wind direction of maximum wind speed
- direction of peak gust speed
- value of maximum wind speed
- count of days with fog

for the location of the building along with the energy consumption of the building in a given year, measured as Site Energy Usage Intensity (Site EUI). Each row of the dataset corresponds to data from a building for a single year. Our task is to predict the Site EUI values for all the buildings given the building and climate data corresponding to the building's location.

The dataset that we have used has two parts: the training dataset and the testing dataset. The training dataset will have one more column corresponding to the site EUI values as compared to the testing dataset where we need to predict that column.



A	B	C	D	E	F
	Facility Type	mean site	standard	max site_	min site_
0	Data_Center	339.74	245.51	829.63	21.74
1	Laboratory	335.13	162.34	782.72	61.819
2	Health_Care_Inpatient	249.29	141.28	887.94	23.934
3	Grocery_store_or_food_market	241.94	81.25	690.56	18.677
4	Food_Service_Restaurant_or_cafeteria	195.59	113.37	455.35	18.449
5	Health_Care_Outpatient_Uncategorized	189.89	128.03	455.52	14.489
6	Health_Care_Uncategorized	183.95	191.48	997.87	2.032
7	Public_Safety_Penitentiary	170.67	101.11	408.85	16.473
8	Public_Assembly_Stadium	157.04	62.107	275.78	80.993
9	Food_Sales	141.45	109.32	855.19	1.362
10	Service_Vehicle_service_repair_shop	140.55	126.04	724.42	1.176
11	Public_Safety_Fire_or_police_station	131.74	56.789	395.82	27.055
12	Nursing_Home	131.17	67.226	661.38	11.657
13	Food_Service_Uncategorized	128.8	61.991	229.66	53.989
14	Public_Assembly_Other	127.17	85.358	648.12	10.28
15	Industrial	124.9	175.89	944.9	1.19
16	Public_Assembly_Entertainment_culture	124.88	85.577	360.19	3.5
17	Service_Uncategorized	119.45	64.287	240.48	17.785
18	Lodging_Other	118.9	70.278	359.81	1.71
19	Commercial_Unknown	118.67	91.985	388.86	11.967
20	Office_Medical_non_diagnostic	116.96	93.284	859.39	16.056
21	Retail_Strip_shopping_mall	109.86	126.86	854.66	1.683
22	Education_College_or_university	109.42	92.332	989.89	1.795
23	Public_Assembly_Library	105.57	64.361	547.66	1.995
24	Lodging_Hotel	105.54	52.619	935.39	5.814
25	Public_Assembly_Recreation	103.83	60.711	317.67	21.142
26	Health_Care_Outpatient_Clinic	103.61	64.349	366.8	9.35
27	Public_Safety_Courthouse	103.18	47.652	233.27	40.689
28	Public_Assembly_Movie_Theater	103.1	35.372	195.32	38.132
29	Retail_Enclosed_mall	101.17	73.933	265.87	3.865
30	Warehouse_Refrigerated	96.284	109.67	813.29	9.486
31	Office_Bank_or_other_financial	90.96	58.121	362.88	21.126
32	Commercial_Other	90.54	94.95	948.76	1.197
33	Mixed_Use_Commercial_and_Residentia	89.852	90.073	903.67	1.379
34	Public_Safety_Uncategorized	84.757	30.932	139.78	9.098
35	Multifamily_Uncategorized	83.831	38.379	993.43	1.003
36	Education_Uncategorized	83.59	45.876	207.63	10.11
37	Office_Mixed_use	82.11	24.562	123.01	32.557
38	Lodging_Dormitory_or_fraternity_sorori	82.071	43.119	351.42	1.097
39	Mixed_Use_Predominantly_Residential	81.798	19.564	109.19	41.673
40	Public_Assembly_Drama_theater	80.917	54.368	233.7	6.214
41	Retail_Uncategorized	80.16	62.348	593.69	2.251
42	Office_Uncategorized	77.041	48.243	970.39	1.003
43	Public_Assembly_Social_meeting	76.102	54.957	194.97	1.097
44	Education_Other_classroom	69.472	34.269	819.08	2.351
45	Parking_Garage	68.229	149.89	975.45	1.199
46	Lodging_Uncategorized	66.595	20.87	89.226	42.029
47	Mixed_Use_Predominantly_Commercial	65.184	71.384	370.29	2.792
48	Public_Assembly_Uncategorized	62.873	52.589	188.45	2.393
49	Education_Preschool_or_daycare	60.974	29.276	203.04	6.083
50	Retail_Vehicle_dealership_showroom	47.007	34.014	187.1	1.4
51	Religious_worship	43.862	31.148	178.15	2.252
52	Service_Drycleaning_or_Laundry	42.113	7.271	54.539	32.209
53	Warehouse_Distribution_or_Shipping_cc	39.599	33.305	233.52	1.31
54	Warehouse_Nonrefrigerated	38.124	48.864	695.99	1.001
55	5plus_Unit_Building	36.719	17.057	159.72	10.314
56	Warehouse_Uncategorized	35.936	28.729	199.78	2.712
57	2to4_Unit_Building	31.878	15.389	149.49	5.079
58	Food_Service_Other	29.237	8.864	51.548	16.2
59	Warehouse_Selfstorage	21.678	16.529	114.3	1.176

Fig1.Table for min, max, mean and standard deviation values for site EUI (click here to see the csv file of this)

2. Dataset Preprocessing

Handling NaN values

There were 6 columns with Nan values (year_built, energy_star_rating, direction_max_wind _speed, direction_peak_wind_speed, max_wind_speed, days_with_fog). The last 4 columns had more than 50% NaN values, so we dropped those columns. Since the data is not a time series data, interpolating the data does not make much sense. We dropped the rows containing NaN value in year_built and filled the NaN values in energy_star_rating with the mean.

Encoding the Categorical Features

We tried two types of encodings - One Hot Encoding and Label Encoding (from sklearn library). After each encoding, we scale the data using StandardScalar (from sklearn library). One hot encoding produced a very sparse matrix. So we did PCA (Principal Component Analysis) using full SVD (Singular Vector Decomposition) on it.

3. Analysis of Data

Analyzing the data helps us get insights into the data and make inferences. The facility type of a building or purpose of a building has a major role in the energy consumption. Thus, we first analyze the data on the basis of the facility type to get an idea of which types of buildings usually have more energy consumption and need to make efficient policies to reduce the same. We have thus calculated the mean, standard deviation, minimum and maximum of the site's Energy Usage Intensity for each facility type to get a better understanding. These values have been reported in Fig1.

We can infer the following from the data:

1. Facilities like Data Centers, Laboratories, and Health units have the most energy consumption amongst all. This might be due to heavy equipment and instruments that are used at these facilities.
2. Grocery stores, food markets, and restaurants also have very high energy usage. This might be due to their storage facilities, where they need to store food items in low temperatures. This is evident from the fact that non-refrigerated warehouses have very less energy usage.
3. . Some facility types have a lot of standard deviation. This is because these facility types have variable sizes according to their scale. For example, industrial buildings can be large scaled and small scaled and thus their energy usage has a lot of variance.
4. Small stores and offices have less energy usage, because they usually have small areas

How can we improve?

Now we have got the facility types that contribute the most in the global energy consumption. Working on these facility types to revise their policies and build efficient strategies can greatly improve the usage of energy. Thus, we further gain insights on how the energy consumption for these facility types have changed over the years. Below we have plotted the trends in the energy usage with time for the facility types with the highest average consumption in Fig2.

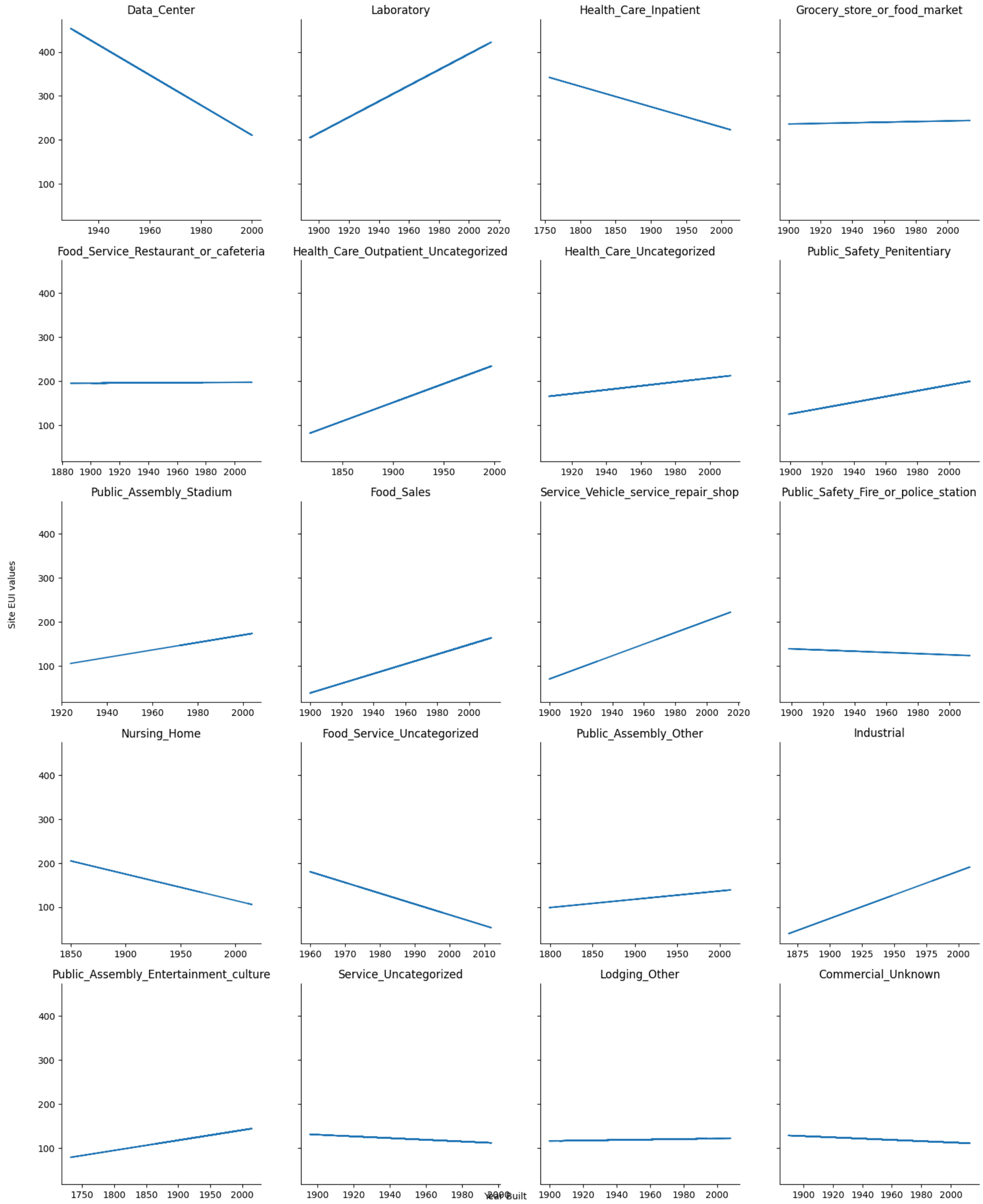


Fig2. Change in Site EUI with time for facility types with most energy consumption.

From the plots, we can make the following inferences:

1. Facility types like Data Center, Health Care Center, Nursing Home, and Uncategorized Food Services have shown a decrease in their energy consumption over the years. This might be a result of optimizations in their energy usage and efficient policies.

2. Facility types like Laboratories, Vehicle Repair Shops, Food Sales, and Industries have shown a significant rise in the consumption of energy. These are the main types where we need to focus and implement efficient strategies to improve energy usage.
3. The above rise might be due to scaling and use of more advanced equipment. But this cannot be continued and we need to optimize energy usage in these facility types to avoid severe consequences.

4. Baseline Implementation

Now we come to the Building Site EUI prediction. Baseline implementation is very important to get a idea of how your model performs. We implemented the simplest regression techniques for comparison, i.e, Linear Regression and Random Forests. We used the inbuilt functions of the scikit-learn library to build the models. Since there are a lot of features in our data, we also implemented LASSO regression[3], which selects features that are more relevant with a L1 penalty in the loss function. We used scikit-learn for this implementation as well.

5. Dense Neural Network

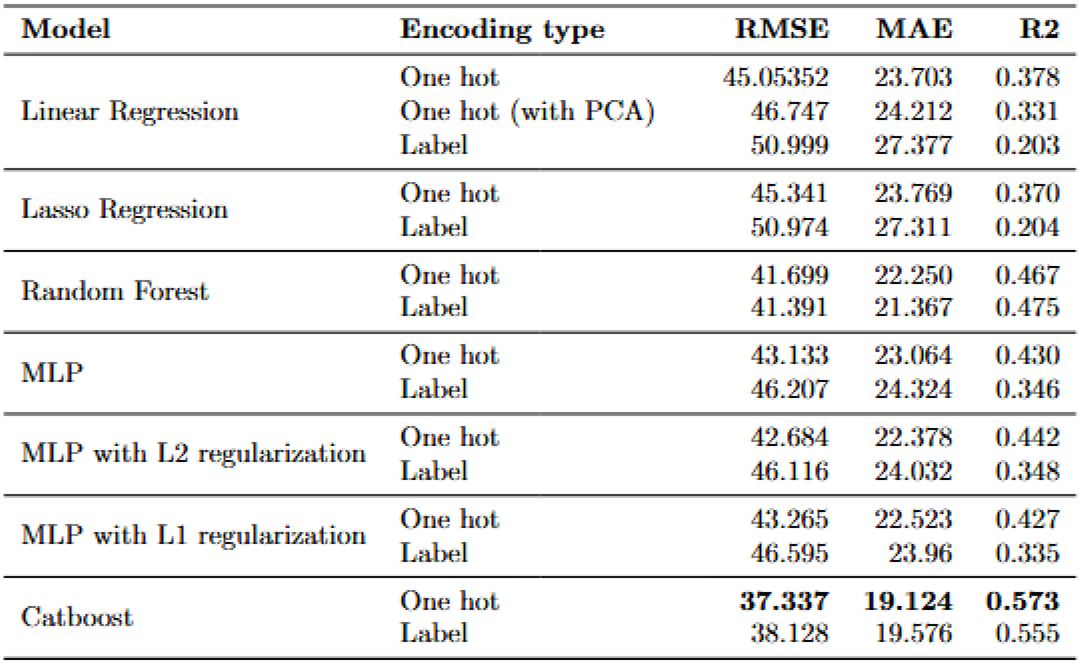
Neural Networks[2] are one of the best Machine Learning techniques that have shown promising results in this domain. We implemented a Dense Neural Network with 6 hidden layers from scratch using the Keras framework. The number of filters in each layer are 1024, 512, 256, 128, 64, and 32 respectively. We used the ReLu activation function on the output of each layer. To avoid overfitting of data, we put a dropout layer with rate 0.2 after each dense layer. We trained the model for 30 epochs with a batch size of 1.

6. CatBoost Regressor

The final model that we implement is the CatBoost Regressor[1]. The reason we chose CatBoost is its ability to handle categorical data, requiring a minimum of categorical feature transformation, opposed to the majority of other machine learning algorithms that cannot handle non-numeric values2 . CatBoost is built upon the theory of Decision Trees and Gradient Boosting. To implement the same we used the catboost library. We trained our model for 5000 epochs with a max depth of 12.

Results and Conclusion

The results for all the Machine Learning techniques for both the preprocessed datasets are summarized in Table below. We have used three metrics for comparison, Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R2).



Model	Encoding type	RMSE	MAE	R2
Linear Regression	One hot	45.05352	23.703	0.378
	One hot (with PCA)	46.747	24.212	0.331
	Label	50.999	27.377	0.203
Lasso Regression	One hot	45.341	23.769	0.370
	Label	50.974	27.311	0.204
Random Forest	One hot	41.699	22.250	0.467
	Label	41.391	21.367	0.475
MLP	One hot	43.133	23.064	0.430
	Label	46.207	24.324	0.346
MLP with L2 regularization	One hot	42.684	22.378	0.442
	Label	46.116	24.032	0.348
MLP with L1 regularization	One hot	43.265	22.523	0.427
	Label	46.595	23.96	0.335
Catboost	One hot	**37.337**	**19.124**	**0.573**
	Label	38.128	19.576	0.555

Table 1: Comparison of different models and encoding types on RMSE, MAE, and R2 metrics.

click here for latex code of this table

We can infer the following from the results:

1. CatBoost gives the best performance among all models with a RMSE score of 37.337. This is because most of the data was categorical and CatBoost excels in fitting categorical data.
2. One hot encoding performs better than label encoding for all of the machine learning models.
3. One hot encoding with PCA is not able to perform as good as without PCA. This maybe because the original data may not have had much noise, thus doing PCA may have removed some necessary information instead of just removing the noise.
4. There is not much difference in the RMSE scores of One Hot encoding and label encoding in CatBoost. This is because the CatBoost model internally selects some of the columns for one-hot encoding.

Summary

In this project, we explored the critical issue of building energy consumption and its impact on climate change. By analyzing a rich dataset of building and weather features, we implemented multiple machine learning models to predict Site Energy Usage Intensity (Site EUI). Through detailed preprocessing, analysis, and modeling, we found that the CatBoost Regressor outperformed other models due to its strength in handling categorical data. Our work highlights the importance of targeted energy efficiency strategies, especially for high-consumption facility types like laboratories and industries, and demonstrates how data-driven approaches can guide sustainable development efforts.