

# **Predictive Modelling of Household Energy Consumption**

*Submitted by*

**JYOTHIRNAYANI GURRALA (YR76819)**

**VENKATA SAI RISHITHA SEELAM (LX79370)**

**DECEMBER 10, 2024**

## Contents

<b>Predictive Modelling of Household Energy Consumption.....</b>	<b>1</b>
Executive Summary .....	3
Key Findings: .....	3
Significance:.....	3
Brief Summary of Methodologies used:.....	4
Introduction.....	4
Research Question.....	5
Background and Context: .....	5
Objectives: .....	6
Data Overview.....	6
Dataset Description:.....	6
Challenges with Data: .....	8
METHODOLOGY: .....	8
Data Analysis:.....	8
Data Cleaning:.....	11
Modelling:.....	13
Assumptions: .....	17
Limitations: .....	18
Overall Results:.....	19
Discussion: .....	20
Significance of key findings and their implication:.....	20
Why These Findings Matter.....	21
Comparison of the Project with Reference Works.....	22
Unexpected Findings: .....	24
Conclusion and Recommendations: .....	24
Future work:.....	24
References:.....	25
Appendix A: GitHub Repository .....	27
Appendix B: Deployment Links.....	27

## **Executive Summary**

This project focuses on predicting household energy consumption to address growing energy demands and environmental concerns. By leveraging a comprehensive dataset of single-family residences in Chicago, the project analyses key factors such as building characteristics, occupancy patterns, and seasonal variations influencing energy usage. Advanced machine learning techniques are employed to forecast electricity and gas consumption, aiming to provide accurate predictions. The insights generated support sustainable urban planning, optimize energy distribution, reduce costs, and encourage environmentally responsible behaviours, making the project scalable and impactful for broader applications in energy management.

### **Key Findings:**

- While analysing the data we came to know the Single Family (subcategory) from Residential exhibits consistent energy change.
- We came to know the seasonal changes may affect the energy consumption.
- The building characteristics also play a major role in influencing changes in energy.
- Among all the machine learning algorithms KNN and Random Forest performed well.

### **Significance:**

Predictive modelling of household energy consumption helps understand how and when energy is used, which is crucial as energy demand increases globally. By analysing past usage data, predictive models can forecast future consumption patterns. For utility companies, these predictions help prevent power shortages, improve grid stability, and better integrate renewable energy sources. For homeowners, it provides insights into when energy use is highest, helping them reduce costs and optimize energy usage. This also encourages the adoption of energy-

efficient practices, such as using appliances at off-peak times or upgrading to energy-saving devices. Overall, it benefits both utilities and consumers by promoting sustainable and cost-effective energy management.

### **Brief Summary of Methodologies used:**

We have used **The Chicago Energy Usage 2010** Dataset. It provides detailed information about energy consumption across Chicago neighbourhoods for the year 2010. It includes data on electricity (kWh) and natural gas (therms) usage, segmented by building types and community areas. We have loaded data into data frame (python) and analysed all the columns. We have drawn few insights from the data. We performed Data cleaning, based on the feature importance columns has been selected. We have divided the data into training and testing data. Performed Standardization on input variables. Trained the data using the machine learning algorithms and calculated the R-squared and MAE values. Made the comparison between the models. In conclusion after applying all Machine learning algorithms, we predict the most accurate model based on R-squared value and MAE value. The predictions made can help utility companies and homeowners manage energy usage more efficiently, contributing to overall energy sustainability.

## **Introduction**

Managing energy consumption in urban areas is a growing challenge, especially in cities like Chicago, where increasing populations and diverse building characteristics make energy demand complex to predict. Seasonal changes, occupancy patterns, and building attributes further influence energy usage, complicating efforts to ensure efficiency and sustainability. Accurate energy forecasting is critical to addressing these challenges, as it supports better resource allocation, cost reduction, and environmental benefits. This project focuses on leveraging advanced machine

learning techniques to predict household energy consumption, providing actionable insights for urban energy management and sustainability.

### **Research Question**

- *How can we leverage advanced machine learning techniques to optimize energy distribution, reduce costs, and promote environmentally responsible behaviors in single-family residences?*

This is important because buildings use a lot of energy and add to pollution. As the population grows and energy needs increase, it's crucial to manage energy use in homes to save money and help the environment. Using Advanced computer programs, we can analyze a lot of data to figure out how to manage energy better. By encouraging eco-friendly habits, this research can help reach sustainability goals and fight climate change. The findings can also help city planners and policymakers create strategies to make homes more energy-efficient and promote sustainability in communities.

### **Background and Context:**

Energy use in cities like Chicago is a big issue because of high demand and the need to be eco-friendly. Predicting how much energy will be used is tough because of things like building size, population, and weather. Using machine learning can make these predictions more accurate. Models like Random Forest and K-Nearest Neighbors have worked well for this. This project is looking at energy use in single-family homes in Chicago. By looking at past energy data, we can make models to predict energy use and figure out what affects it the most. The aim is to help plan cities in a sustainable way, create better energy policies, and promote habits that save energy. This can help reduce Chicago's carbon footprint and reach environmental goals.

## Objectives:

1. **Forecast Energy Consumption:** Create predictive models to estimate monthly and total electricity (kWh) and gas (therms) usage for single-family residential properties in Chicago. Evaluate the accuracy and robustness of these models.
2. **Identify Key Drivers:** Analyze how factors such as building age, size, population, and occupancy influence energy consumption.
3. **Support Sustainable Development:** Provide insights to guide energy policies, urban planning, and community outreach programs to reduce energy consumption and improve efficiency.

This project helps the city with its energy problems and can also be used in other cities around the world to save energy, make money, and help the environment.

## Data Overview

### Dataset Description:

We have considered **The Chicago Energy Usage 2010** Dataset. It provides detailed information about energy consumption across Chicago neighbourhoods for the year 2010. It includes data on electricity (kWh) and natural gas (therms) usage, segmented by building types and community areas.

### Key Characteristics of the Dataset:

- Source: Public utility data from Chicago for the year 2010.
- Size: 67,051 rows  $\times$  73 columns.
- Time Period: Covers monthly electricity and gas consumption for 12 months.

- Scope: Focused on residential energy usage across different building types, including single-family homes, multi-unit dwellings, and commercial structures.
- Below are the few important variables in data set.

Column Name	Description
<b>TOTAL THERMS</b>	Total Gas consumption in therms in 2010.
<b>TOTAL KWH</b>	Total electricity consumed in 2010 in kWh.
<b>GAS ACCOUNTS</b>	Number of accounts with THERM information
<b>ELECTRICITY ACCOUNTS</b>	Number of accounts with kilowatt hour information
<b>TOTAL POPULATION</b>	Total population from Census 2010 report
<b>OCCUPIED UNITS</b>	Number of housing units that are occupied
<b>RENTER-OCCUPIED</b>	Number of housing units that are renter occupied
<b>RENTER-OCCUPIED HOUSING PERCENTAGE</b>	Percentage of occupied housing units that are renters from 'Census report
<b>ZERO KWH ACCOUNTS</b>	Number of accounts with 0 kilowatt hours amounts for 12 months in 2010
<b>KWH TOTAL SQFT</b>	Total square footage associated with the electric energy usage
<b>THERMS TOTAL SQFT</b>	Total square footage associated with the natural gas energy usage for Kilowatt Hours
<b>TOTAL UNITS</b>	Total number of housing units
<b>AVERAGE STORIES</b>	Average number of stories
<b>AVERAGE BUILDING AGE</b>	Average age of Building
<b>AVERAGE HOUSE SIZE</b>	Average household size from Census 2010 report
<b>OCCUPIED UNITS PERCENTAGE</b>	Occupied units percentage.
<b>OCCUPIED HOUSING UNITS</b>	Number of occupied housing units from 'Census 2010 report

<b>BUILDING_TYPE</b>	Building Type: Residential, Commercial, Industrial
<b>BUILDING_SUBTYPE</b>	Building Sub-Type (6): Single Family, Multi <7, Multi 7+, Commercial, Industrial, Municipal.

### Challenges with Data:

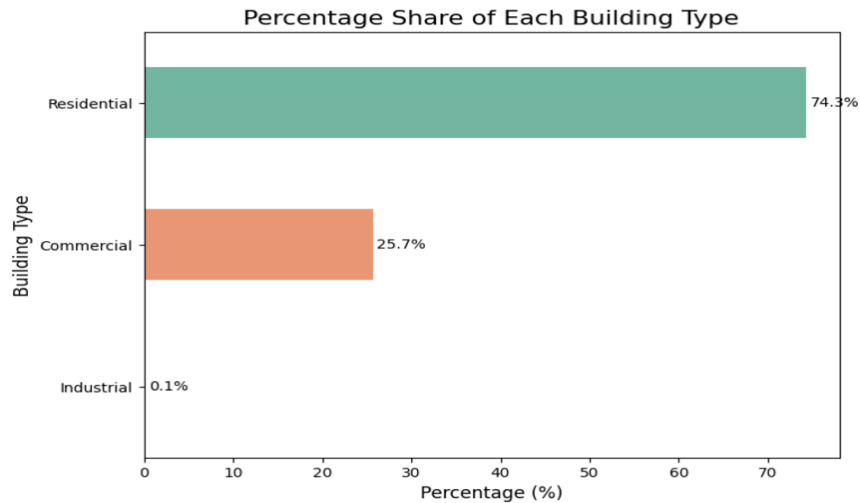
- Presence of missing values in the data
- There are 73 columns in the data. So required analysis should be made to choose the features that affect the target variable.
- Inconsistent naming in some columns, such as TERM APRIL 2010 instead of THERM APRIL 2010. The column has been renamed manually.

### METHODOLOGY:

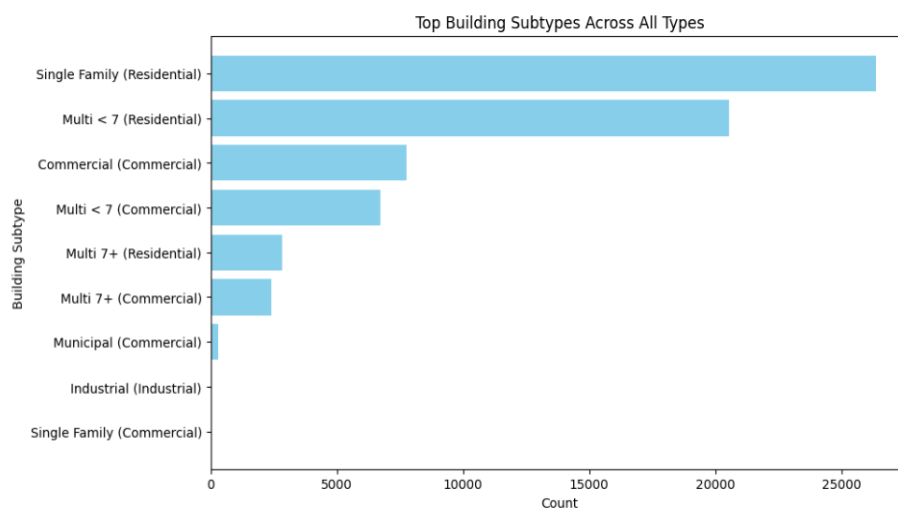
#### Data Analysis:

- There are three building types available in our dataset. **Residential (49,747)**, **Commercial (17,185)** and **Industrial (42)**. Among the 3 building types Residential holds more proportion. Building subtypes are divided into (**Multi < 7, Single Family, Commercial, Multi 7+, Municipal, Industrial**)

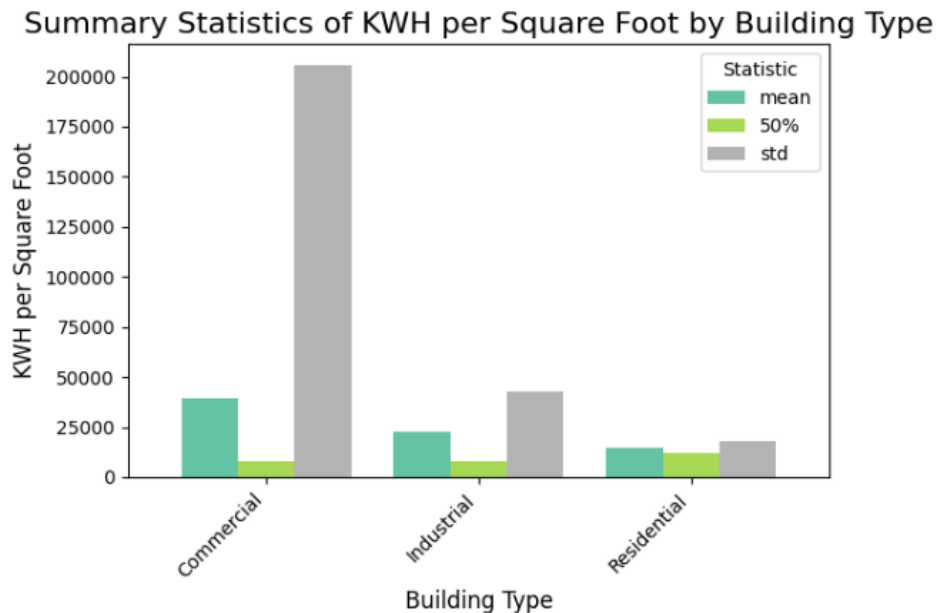




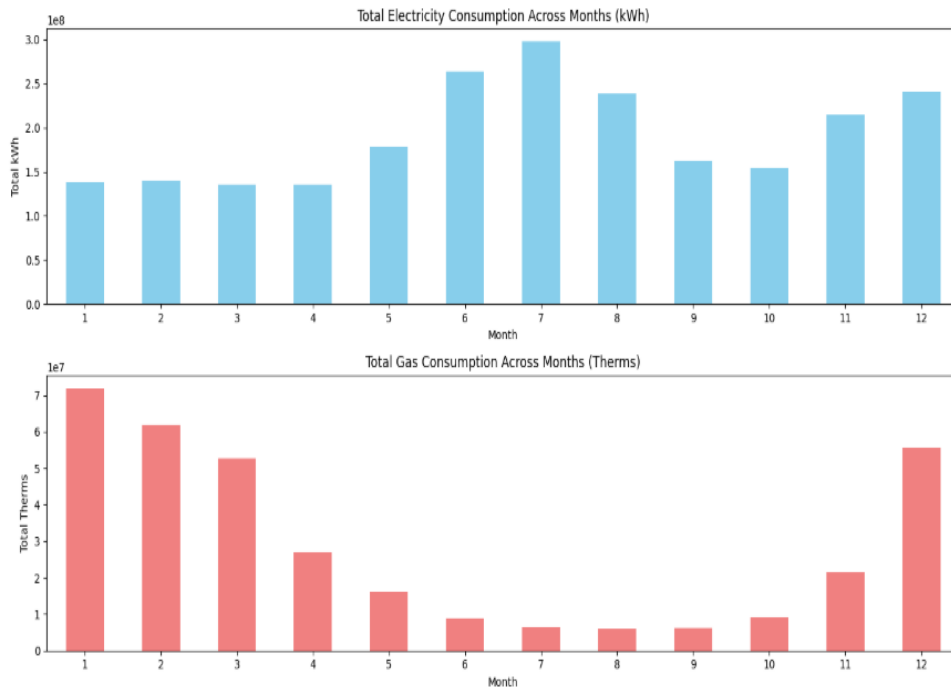
- We have grouped the data based on the building type and we listed the subtypes in descending order. We can see in the below figure.



- Based on each building category we calculated the energy variation over square feet. We observed both Industrial and Commercial building types showed more variation of electricity per square foot. The energy patterns are not consistent to apply ML algorithms and train the model. The residential Single family showed consistent variations in energy. So we decided to focus more on Residential Single Family.



- We checked the dependency of other feature variables on the target. Community Name, Building age doesn't strongly affect the target variable.
- We also analysed the Renter occupied units versus total population.
- We have also noticed the monthly energy usage trends. The higher energy consumption in July and June is due to summer temperatures, increased usage of air conditioning and cooling systems.
- Similarly, heightened energy demand in December and November is influenced by the seasonal need for heating during colder months, coupled with additional usage for festive decorations and increased building occupancy during holiday celebrations. During winter months (January to March) and especially in December, gas consumption peaks due to heightened heating needs driven by cold temperatures.



### Data Cleaning:


- We have done our analysis on various features, and we observed that energy pattern is consistent for Residential – Single Family comparing to other sub-types.
- Based on the analysis, residential single-family data plays a dominant role, making it ideal for insights into household energy use.
- Excluding commercial and industrial buildings is justified, as these categories have unique energy demands and minimal single-family presence (only 1 for commercial and 42 for industrial). Residential data contains the largest single-family count (26,365), making it central to understanding urban household energy consumption.
- We have changed dataset from a wide format (where each month has its own column) into a long format (where each row represents one month's energy usage for a specific location).

We have removed unnecessary summary rows like 'MEAN 2010' and cleaned up the month labels by removing the year (e.g., changing 'January 2010' to 'January').

- We have converted month names (e.g., "JANUARY") to their corresponding numeric values (e.g., 1). We have removed outliers from data and replaced nulls with the mean of respective column.
- Both Electric accounts and Gas accounts contain 'Less than 4' entry. This has been removed.

data\_0

COMMUNITY AREA NAME	CENSUS BLOCK	BUILDING TYPE	BUILDING_SUBTYPE	KWH JANUARY 2010	KWH FEBRUARY 2010	KWH MARCH 2010	KWH APRIL 2010	KWH MAY 2010	KWH JUNE 2010	...
Hyde Park	1.703141e+14	Residential	Single Family	NaN	NaN	NaN	NaN	NaN	NaN	...
Lakeview	1.703106e+14	Residential	Single Family	16620.0	13420.0	8570.0	6124.0	5972.0	7081.0	...
Logan Square	1.703183e+14	Residential	Single Family	NaN	NaN	NaN	NaN	NaN	NaN	...
New City	1.703184e+14	Residential	Single Family	188.0	482.0	322.0	263.0	184.0	837.0	...
North Center	1.703105e+14	Residential	Single Family	1602.0	1273.0	1186.0	1068.0	1496.0	2361.0	...



	COMMUNITY AREA NAME	CENSUS BLOCK	Month	AVERAGE BUILDING AGE	AVERAGE HOUSESIZE	AVERAGE STORIES	Age Category	BUILDING TYPE	BUILDING_SUBTYPE	ELECTRICITY ACCOUNTS	...	THERMS SQFT MINIMUM 2010	THERMS SQFT STANDARD DEVIATION 2010
0	Hyde Park	1.703141e+14	1	0.0	1.96	1.0	NaN	Residential	Single Family	NaN	...	3872.0	355.52
1	Hyde Park	1.703141e+14	2	0.0	1.96	1.0	NaN	Residential	Single Family	NaN	...	3872.0	355.52
2	Hyde Park	1.703141e+14	3	0.0	1.96	1.0	NaN	Residential	Single Family	NaN	...	3872.0	355.52
3	Hyde Park	1.703141e+14	4	0.0	1.96	1.0	NaN	Residential	Single Family	NaN	...	3872.0	355.52
4	Hyde Park	1.703141e+14	5	0.0	1.96	1.0	NaN	Residential	Single Family	NaN	...	3872.0	355.52
5	Hyde Park	1.703141e+14	6	0.0	1.96	1.0	NaN	Residential	Single Family	NaN	...	3872.0	355.52
6	Hyde Park	1.703141e+14	7	0.0	1.96	1.0	NaN	Residential	Single Family	NaN	...	3872.0	355.52
7	Hyde Park	1.703141e+14	8	0.0	1.96	1.0	NaN	Residential	Single Family	NaN	...	3872.0	355.52
8	Hyde Park	1.703141e+14	9	0.0	1.96	1.0	NaN	Residential	Single Family	NaN	...	3872.0	355.52
9	Hyde Park	1.703141e+14	10	0.0	1.96	1.0	NaN	Residential	Single Family	NaN	...	3872.0	355.52
10	Hyde Park	1.703141e+14	11	0.0	1.96	1.0	NaN	Residential	Single Family	NaN	...	3872.0	355.52
11	Hyde Park	1.703141e+14	12	0.0	1.96	1.0	NaN	Residential	Single Family	NaN	...	3872.0	355.52
15	Lakeview	1.703106e+14	1	0.0	1.30	1.0	NaN	Residential	Single Family	8	...	NaN	NaN
16	Lakeview	1.703106e+14	2	0.0	1.30	1.0	NaN	Residential	Single Family	8	...	NaN	NaN
17	Lakeview	1.703106e+14	3	0.0	1.30	1.0	NaN	Residential	Single Family	8	...	NaN	NaN

- We have calculated the correlation of the features with target variables.

- Few features are highly co-related with others this may cause difficulty to distinguish their effect on the target variables. So we decided to 'GAS ACCOUNTS' and 'THERMS TOTAL SQFT'. This reduces redundancy, improves model interpretability, and prevents overfitting while maintaining predictive power.

## **Modelling:**

### **Data Standardization and Feature Selection:**

- We have chosen these columns for data analysis:  
  
'TOTAL KWH', 'TOTAL THERMS','MONTHLY KWH', 'MONTHLY THERMS',  
'Month', 'ELECTRICITY ACCOUNTS', 'GAS ACCOUNTS','TOTAL POPULATION',  
'AVERAGE STORIES', 'AVERAGE BUILDING AGE','THERMS TOTAL SQFT',  
'TOTAL UNITS', 'ZERO KWH ACCOUNTS', 'OCCUPIED UNITS', 'OCCUPIED UNITS  
PERCENTAGE', 'AVERAGE HOUSESIZE', 'RENTER-OCCUPIED HOUSING UNITS',  
'RENTER-OCCUPIED HOUSING PERCENTAGE', 'KWH TOTAL SQFT'
- The target variables are 'TOTAL KWH', 'TOTAL THERMS', 'MONTHLY KWH',  
'MONTHLY THERMS'.
- **OneHotEncoder** is applied to categorical columns like Month, converting them into binary features. **StandardScaler** is used to standardize numerical columns (e.g., energy and housing metrics), scaling them to have a mean of 0 and standard deviation of 1. The target variables are not standardized.
- We have divided the data into (75% training and 25% testing data). Since all the input and target variables are continuous, we use Regression Models to predict the target variables.
- We have applied below Machine Learning Algorithms.

## 1. Linear Regression:

Since our target variables and dependent variable are continuous, we predict a continuous outcome- we chose Linear Regression model. Below are the train and test predictions. The reason why we see the MAE in big numbers is the target variables are not standardized. The model performs well in predicting TOTAL KWH and TOTAL THERMS, as indicated by high R-squared values ( $\sim 0.75$ ) and consistent MAE across train and test sets, showing good generalization. However, predictions for MONTHLY KWH and MONTHLY THERMS are less accurate, with lower R-squared values ( $\sim 0.56$ – $0.63$ ), likely due to higher variability in monthly data.

	Model	Metric	KWH Train Set	THERMS Train Set	Monthly KWH Train Set	Monthly THERMS Train Set	KWH Test Set	THERMS Test Set	Monthly KWH Test Set	Monthly THERMS Test Set
0	Linear Regression	R-squared	0.7548	0.7467	0.5665	0.6298	0.7523	0.7441	0.5630	0.6272
1	Linear Regression	MAE	16214.9560	2370.6103	2118.9463	384.6511	16284.4672	2378.4241	2125.8786	387.6959

## 2. Polynomial Regression:

Polynomial Regression (degree=3) outperforms Linear Regression in terms of both R-squared and Mean Absolute Error (MAE) across all target variables. The model shows a better fit to the data with higher R-squared values, indicating improved prediction accuracy. Polynomial Regression enhances the model's ability to capture complex relationships in the data, resulting in better overall performance. The model shows a significant improvement over Linear Regression in terms of R-squared and MAE.

	Model	Metric	KWH Train Set	THERMS Train Set	Monthly KWH Train Set	Monthly THERMS Train Set	KWH Test Set	THERMS Test Set	Monthly KWH Test Set	Monthly THERMS Test Set
0	Polynomial Regression	R-squared	0.7757	0.7783	0.6309	0.7366	0.7706	0.7722	0.6260	0.7333
1	Polynomial Regression	MAE	15604.5436	2220.3542	1959.0420	298.0540	15776.5489	2248.5220	1970.9651	301.5379

3. **Decision Tree:** The R-squared values for the train set are high across all target variables. On the test set, the R-squared values are slightly lower but still reasonably high, indicating that the model generalizes well. The MAE values for the train set are relatively low, indicating good prediction accuracy on the training data. On the test set, the MAE values increase slightly, as expected, showing that the model's predictions on unseen data are a bit less accurate.

**Summary:** The Decision Tree model performs well overall, with good fit (R-squared values) and low prediction errors (MAE) for both training and test sets.

	Model	Metric	KWH Train Set	THERMS Train Set	Monthly KWH Train Set	Monthly THERMS Train Set	KWH Test Set	THERMS Test Set	Monthly KWH Test Set	Monthly THERMS Test Set
0	Decision Tree	R-squared	0.8278	0.8405	0.6400	0.7264	0.8168	0.8278	0.6099	0.6895
1	Decision Tree	MAE	13900.8949	1936.9773	1949.6918	319.1322	14344.0572	2004.3376	2020.7003	339.4855

4. **Random Forest:** The R-squared values for the train set are quite high, indicating a good fit of the model to the training data. On the test set, the R-squared values drop slightly but remain strong, showing good generalization. The MAE values for the train set are relatively low, meaning the model has a small average error in predictions. On the test set, the MAE values increase slightly, which is typical for most models when applied to unseen data.

**Summary:** The Random Forest model performs well, with high R-squared values (indicating a good fit) and low MAE values (indicating accurate predictions) for both

the training and test sets. The model shows some reduction in performance on the test set, which is expected when generalizing to new data, but the decrease is minimal, suggesting good generalization.

	Model	Metric	KWH Train Set	THERMS Train Set	Monthly KWH Train Set	Monthly THERMS Train Set	KWH Test Set	THERMS Test Set	Monthly KWH Test Set	Monthly THERMS Test Set
0	Random Forest	R-squared	0.8531	0.8634	0.6547	0.7415	0.8421	0.8524	0.6315	0.7077
1	Random Forest	MAE	13123.5077	1826.7508	1920.3586	313.4253	13573.5311	1897.0038	1976.8980	333.8879

5. **Gradient Boosting:** The R-squared values for the train set are moderate. For the test set, the R-squared values are slightly lower but still reflect reasonable model performance. The MAE values for the train set show higher errors compared to the Random Forest model, indicating that the model experiences some error increase when applied to new, unseen data.

**Summary:** Gradient Boosting shows decent R-squared values, but the performance is slightly lower compared to models like Random Forest in both training and test sets. The MAE is relatively high, suggesting the model may not be as precise in its predictions compared to other models, though still reasonably accurate. While there is some drop in performance from training to test set (as expected), the Gradient Boosting model demonstrates acceptable predictive accuracy.

	Model	Metric	KWH Train Set	THERMS Train Set	Monthly KWH Train Set	Monthly THERMS Train Set	KWH Test Set	THERMS Test Set	Monthly KWH Test Set	Monthly THERMS Test Set
0	Gradient Boosting	R-squared	0.7838	0.7977	0.6151	0.7021	0.7747	0.7840	0.6058	0.6972
1	Gradient Boosting	MAE	15599.1108	2172.9533	2021.4490	327.8389	15719.8177	2203.1258	2025.7173	332.0542

6. **Extreme Gradient Boosting:** On the training set, the R-squared values are strong. On the test set, the R-squared values drop slightly but remain decent. These values show



that the model can generalize reasonably well, though some performance is lost when moving from training to test data. The MAE values on the training set show that the model is relatively accurate. For the test set, MAE values increase somewhat, with KWH having an MAE of 15,660.74, indicating a slight drop in prediction accuracy on unseen data.

**Summary:** The XG Boosting model performs well on both train and test sets, with R-squared values that suggest good model fit, particularly for THERMS. The MAE values show a slight increase in error when applying the model to the test data.

	Model	Metric	KWH Train Set	THERMS Train Set	Monthly KWH Train Set	Monthly THERMS Train Set	KWH Test Set	THERMS Test Set	Monthly KWH Test Set	Monthly THERMS Test Set
0	XG Boosting	R-squared	0.8418	0.8611	0.6703	0.7735	0.7790	0.7747	0.5998	0.7168
1	XG Boosting	MAE	13733.4056	1856.5748	1868.9769	268.5468	15660.7397	2274.0831	1995.8674	307.7796

7. **KNN:** The KNN model performs well on the training set with high R-squared and low MAE but shows reduced generalization on the test set, particularly for monthly energy predictions.

	Model	Metric	KWH Train Set	THERMS Train Set	Monthly KWH Train Set	Monthly THERMS Train Set	KWH Test Set	THERMS Test Set	Monthly KWH Test Set	Monthly THERMS Test Set
0	KNN	R-squared	0.8890	0.8958	0.7257	0.7595	0.8365	0.8486	0.5878	0.6333
1	KNN	MAE	10572.3830	1474.0033	1681.1788	293.0838	12811.2936	1776.4389	2071.2290	363.9823

### Assumptions:

- **Data Represents Reality:** The dataset is assumed to reflect actual energy usage patterns for 2010 in Chicago.
- **Stable Patterns:** Energy usage is assumed to stay consistent over time for making predictions.

- **Independent Features:** The model assumes that input features (like building age or size) don't strongly depend on each other.
- **No Sudden Behavior Changes:** People's energy usage habits are assumed to be consistent throughout the dataset.
- **External Factors Ignored:** Things like changes in energy pricing or government policies are not considered.

#### **Limitations:**

- **Old Data:** The dataset is from 2010, and energy usage patterns might have changed over the years.
- **Chicago-Specific:** The model is trained for Chicago, so it might not work well for other cities with different climates or lifestyles.
- **Missing Details:** The model doesn't include factors like weather or economic conditions, which could impact energy usage.
- **Simplistic Models:** Some algorithms assume simple relationships (like Linear Regression assumes a straight-line relationship), which may not always match reality.
- **Limited Scope:** The model focuses on single-family homes and may not work well for other types of buildings like offices or factories.
- **Unexpected Changes:** Sudden shifts in energy usage, like during heatwaves or power outages, are not accounted for.

Overall Results:

Results after applying various ML Algorithms

#summary of accuracy % for different models  
Results\_final

	Model	Metric	KWH Train Set	THERMS Train Set	Monthly KWH Train Set	Monthly THERMS Train Set	KWH Test Set	THERMS Test Set	Monthly KWH Test Set	Monthly THERMS Test Set	
0	Linear Regression	R-squared	0.7548	0.7467	0.5665	0.6298	0.7523	0.7441	0.5630	0.6272	
1	Linear Regression	MAE	16214.9560	2370.6103	2118.9463	384.6511	16284.4672	2378.4241	2125.8786	387.6959	
2	Polynomial Regression	R-squared	0.7757	0.7783	0.6309	0.7366	0.7706	0.7722	0.6260	0.7333	
3	Polynomial Regression	MAE	15604.5436	2220.3542	1959.0420	298.0540	15776.5489	2248.5220	1970.9651	301.5379	
4	Decision Tree	R-squared	0.8278	0.8405	0.6400	0.7264	0.8168	0.8278	0.6099	0.6895	
5	Decision Tree	MAE	13900.8949	1936.9773	1949.6918	319.1322	14344.0572	2004.3376	2020.7003	339.4855	
6	Randomn Forest	R-squared	0.8531	0.8634	0.6547	0.7415	0.8421	0.8524	0.6315	0.7077	
7	Randomn Forest	MAE	13123.5077	1826.7508	1920.3586	313.4253	13573.5311	1897.0038	1976.8980	333.8879	
8	Gradient Boosting	R-squared	0.7838	0.7977	0.6151	0.7021	0.7747	0.7840	0.6058	0.6972	
9	Gradient Boosting	MAE	15599.1108	2172.9533	2021.4490	327.8389	15719.8177	2203.1258	2025.7173	332.0542	
10	XG Boosting	R-squared	0.8418	0.8611	0.6703	0.7735	0.7790	0.7747	0.5998	0.7168	
11	XG Boosting	MAE	13733.4056	1856.5748	1868.9769	268.5468	15660.7397	2274.0831	1995.8674	307.7796	
12	KNN	R-squared	0.8890	0.8958	0.7257	0.7595	0.8365	0.8486	0.5878	0.6333	
13	KNN	MAE	10572.3830	1474.0033	1681.1788	293.0838	12811.2936	1776.4389	2071.2290	363.9823	

Snippet of the US Household Energy Prediction Model Web Application:

We have calculated the our target variables (Monthly gas and power consumption), (total power and gas consumption)

Household Energy Prediction System

Energy prediction based on area sqft

Energy prediction Trend

Energy prediction based on area(sqft)

Select a month:

January

OCCUPIED HOUSING UNITS

6

AVERAGE STORIES

1

KWH TOTAL SQFT

14533

OCCUPIED UNITS PERCENTAGE

0.81

TOTAL POPULATION

48

ELECTRICITY ACCOUNTS

14

AVERAGE BUILDING AGE

88

ZERO KWH ACCOUNTS

2

OCCUPIED HOUSING PERCENTAGE

0.35

AVERAGE HOUSESIZE

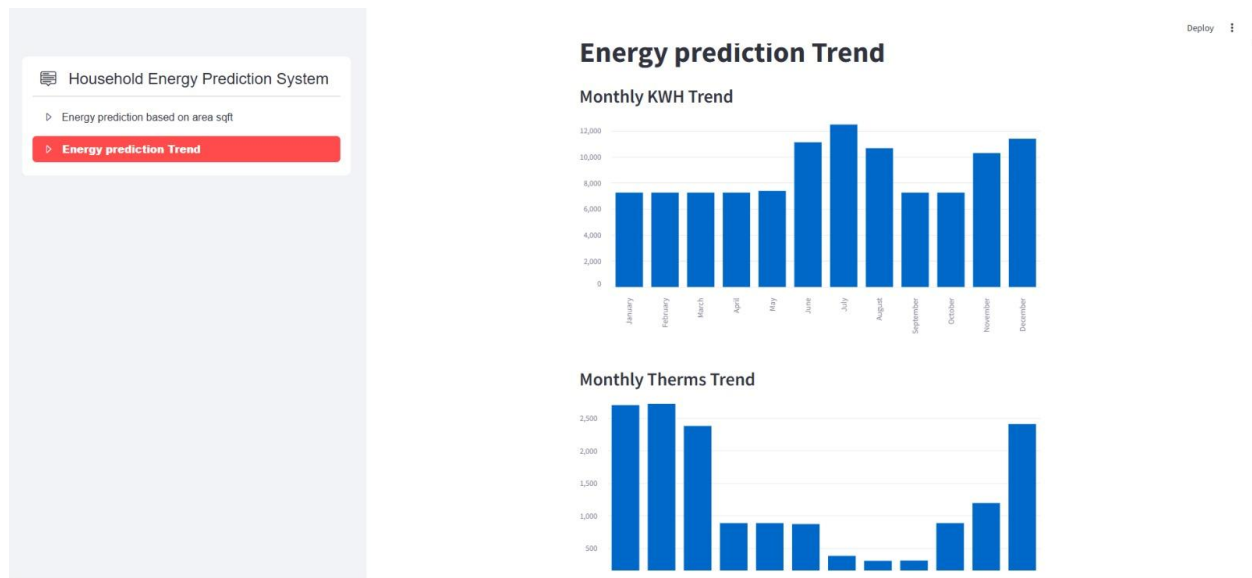
2

Show Household Energy Consumption

ANNUAL\_KWH: 105140.68 & ANNUAL\_THERMS: 18740.48

January: MONTHLY\_KWH: 7245.63 & MONTHLY\_THERMS: 2698.00

Reset



## Discussion:

### Significance of key findings and their implication:

#### 1. Consistent Energy Use in Single-Family Homes

- **What we found:** Single-family homes show steady patterns in energy consumption.
- **Why it matters:** This makes them a predictable group for studying energy use and creating strategies.
- **Impact:** Utility companies and city planners can design programs specifically for these homes, like energy-saving campaigns or better energy distribution plans.

#### 2. Seasonal Changes Affect Energy Use

- **What we found:** Energy consumption increases or decreases with the seasons, especially due to heating and cooling needs.
- **Why it matters:** It helps energy providers understand when demand will peak, like in summer or winter.

- **Impact:** This allows them to manage energy grids better during high-demand seasons and encourage people to save energy during these times.

### 3. **Building Characteristics Are Important**

- **What we found:** Factors like the size of the building, its age, and how many people live there affect energy consumption.
- **Why it matters:** Buildings with certain features may use more energy than others, highlighting areas for improvement.
- **Impact:** Planners and architects can focus on making older or larger buildings more energy-efficient, helping reduce overall energy use.

### 4. **KNN and Random Forest Are Best Models**

- **What we found:** Among all the models tested, KNN and Random Forest were the most accurate in predicting energy use.
- **Why it matters:** These models are good at understanding patterns in the data and making reliable predictions.
- **Impact:** These models can be used in other cities or regions to predict energy usage and help make better decisions for energy planning.

### **Why These Findings Matter**

These findings can help reduce energy waste, lower costs for homeowners, and support sustainable development. Utility companies can plan energy supply more effectively, and residents can save money by learning about better energy-saving practices. Overall, these insights can lead to smarter energy management, benefiting both the community and the environment.

## Comparison of the Project with Reference Works

1. S.S.K. Kwok et al. (2011): A study of the importance of occupancy to building cooling load in prediction by intelligent approach

- **Focus:** This study emphasizes the role of building occupancy in predicting cooling loads, highlighting how occupancy significantly impacts energy consumption.
- **Comparison:**
  - **Similarities:** Both studies acknowledge that building characteristics and usage patterns, including occupancy, influence energy consumption. In our project, factors like building size, occupancy percentage, and housing type were also considered important variables.
  - **Differences:** Our project includes seasonal variations and focuses on total electricity and natural gas usage, while Kwok et al. primarily concentrate on cooling load prediction using intelligent models.
  - **Implications:** The inclusion of occupancy in energy models is crucial, and our findings align with Kwok et al.'s work in recognizing its significance for accurate predictions.

2. H. Zhao et al. (2012): A review on the prediction of building energy consumption

- **Focus:** This review outlines various techniques, such as statistical, machine learning, and hybrid methods, for predicting building energy consumption and emphasizes the importance of reliable models.
- **Comparison:**

- **Similarities:** Both works utilize machine learning models like Random Forest and Support Vector Machines for prediction. Our project also explores seasonal and structural impacts on energy usage, consistent with Zhao et al.'s focus on external factors affecting energy consumption.
- **Differences:** Zhao et al. provide a broader overview of methodologies for diverse building types, while our project focuses on single-family residential buildings in Chicago.
- **Implications:** Our project validates Zhao et al.'s recommendation of machine learning techniques by demonstrating the effectiveness of algorithms like KNN and Random Forest in forecasting energy usage.

3. A.S. Ahmad et al. (2014): A review on applications of ANN and SVM for building electrical energy consumption forecasting

- **Focus:** This study examines the application of Artificial Neural Networks (ANN) and Support Vector Machines (SVM) in predicting electrical energy consumption.
- **Comparison:**
  - **Similarities:** Our project also incorporates SVM as one of the models for predicting energy consumption. Both studies highlight the value of advanced algorithms in improving prediction accuracy.
  - **Differences:** Ahmad et al. focus on ANN and SVM applications specifically, while our project evaluates multiple algorithms (e.g., KNN, Random Forest, Gradient Boosting) to identify the best-performing model.

- **Implications:** Ahmad et al.'s work supports the scalability of advanced algorithms for energy prediction, and our project demonstrates similar applicability in residential energy forecasting.

### **Unexpected Findings:**

- The Community Area does not show strong impact on the target variables

### **Conclusion and Recommendations:**

1. Energy-saving measures are most suited for single-family homes in residential regions because of their steady energy consumption.
2. Due to heating and cooling requirements, energy demand varies with the seasons, rising in the winter and summer.
3. The KNN and Random Forest models gave the greatest predictions for energy use, making them trustworthy instruments for energy planning and forecasting.
4. Data from the dataset indicates where energy-efficient appliances, improved insulation, and renewable energy sources should be promoted.
5. By using predictive models, utility companies may better manage their resources, cutting waste and enhancing grid stability.
6. These results highlight the significance of focused interventions, community education, and sophisticated prediction techniques in attaining sustainability

### **Future work:**

#### **Including Real-Time Data:**

- To increase the precision and promptness of forecasts, future research could use real-time energy consumption data.



**Extending the Dataset:**

- To analyze trends and improve the model's scalability across other geographic locations, we can use data from several years or other cities.

**Including Weather Data:**

- The model's capacity to take seasonal variations into account may be enhanced by incorporating specific weather variables like temperature, humidity, and wind.

**Examining Complex Algorithms:**

- We can try using deep learning methods to capture patterns of time-dependent energy use, such as transformers or Long Short-Term Memory (LSTM) networks.

**Segmented study:**

- To create specialized energy-saving plans for each category, expand the study to include other building subtypes (such as commercial or industrial).

**Behavioral Insights:**

- Examine how patterns of appliance use or household behaviors affect energy use and offer solutions.

**References:**

1. S.S.K. Kwok et al.

[A study of the importance of occupancy to building cooling load in prediction by intelligent approach](#)

Energy Convers Manag (2011)

2. H. Zhao et al.

[A review on the prediction of building energy consumption](#)

Renew Sustain Energy Rev (2012)

3. A.S. Ahmad et al.

[A review on applications of ANN and SVM for building electrical energy consumption forecasting](#)

Renew Sustain Energy Rev (2014)

4. Data Source - City of Chicago Open Data Portal

<https://data.cityofchicago.org/Environment-Sustainable-Development/Energy-Usage-2010/8yq3-m6wp/data>

5. Tools/Software used:

[Google Colab - https://colab.research.google.com/](https://colab.research.google.com/)

## Appendix A: GitHub Repository

### Repository Overview

The overview of the repository structure and its contents:

Data/: Data Set of Energy Usage.

Notebook/: Jupyter notebook for Exploratory Data Analysis, Visualization and Streamlit API files.

Documents/: Presentation Slides and Report for the Project Predictive modelling of Household Energy Consumption.

README.md: Overview of the Project.

GitHub Repository Link

<https://github.com/jyothirayani/UMBC---DATA606-Capstone.git>

## Appendix B: Deployment Links

### Model Deployment



### Streamlit Application

Description: "The frontend is designed to allow users to input features such as Month, Occupied Housing units, Average stories, KWH Total SQFT, Occupied Units percentage, Total Population, Electricity Accounts, Average Building Age, Zero KWH Accounts, Occupied Housing Percentage, Average House Size to predict household energy consumption."