**IST 687 - INTRODUCTION TO DATA SCIENCE**

**FINAL PROJECT REPORT**

**Team Members**

**Bhavesh, Karan, Prathmesh, Pratik, Praveen**

**Prof. Christopher Dunham**

**&**

**Instructor: Akit Kumar**

Contents

[1. INTRODUCTION 4](#_Toc165501975)

[a. Project Background 4](#_Toc165501976)

[b. Project Commission 4](#_Toc165501977)

[2. Project Objective 5](#_Toc165501978)

[a. Develop a Predictive Model: 5](#_Toc165501979)

[b. Understand Key Drivers of Energy Usage: 5](#_Toc165501980)

[3. Data Sources and Preparation 6](#_Toc165501981)

[a. Static House Data: 6](#_Toc165501982)

[b. Energy Usage Data: 6](#_Toc165501983)

[c. Weather Data: 6](#_Toc165501984)

[4. Methodology 7](#_Toc165501985)

[a. Data Collection: 7](#_Toc165501986)

[I. Static-House Data 7](#_Toc165501987)

[II. Weather Data 8](#_Toc165501988)

[III. Energy Usage Data 8](#_Toc165501989)

[b. Data Analysis (EDA): 9](#_Toc165501990)

[I. Distribution of Square Footage 9](#_Toc165501991)

[II. Square Footage VS. Bedrooms by Number of stories 10](#_Toc165501992)

[III. Wind Rose 11](#_Toc165501993)

[IV. Distribution of HVAC Heating type 12](#_Toc165501994)

[V. Distribution of HVAC Cooling types 13](#_Toc165501995)

[VI. Energy Consumption Over time 14](#_Toc165501996)

[VII. Effect of Temperature on heating and cooling: 15](#_Toc165501997)

[c. Predictive Modeling: 16](#_Toc165501998)

[d. Model Evaluation and Refinement: 17](#_Toc165501999)

[e. Shiny App 18](#_Toc165502000)

[5. Key Findings 20](#_Toc165502001)

[6. Recommendations 21](#_Toc165502002)

[7. Conclusion 22](#_Toc165502003)

# INTRODUCTION

## Project Background

The energy consumption project was initiated due to growing concerns about the impacts of global warming on electricity demand, particularly during the hot summer months of July in South Carolina. This period is marked by significantly increased cooling needs as temperatures rise, leading to higher energy consumption in residential areas. An energy company, tasked with supplying electricity to these areas, faces challenges in managing the peak demand to avoid overloading the grid, which could result in blackouts. The company's concern is not only about maintaining a stable supply but also about finding sustainable ways to manage energy consumption without having to expand its current infrastructure through costly developments like new power plants.

## Project Commission

The project was commissioned by the energy company as a proactive measure to address these challenges. The goal was to leverage data-driven insights to better predict and manage periods of high energy demand. By understanding the specific factors that drive energy usage, the company could implement targeted strategies to encourage energy conservation among its customers, thus reducing the overall load on the grid during peak times.

# Project Objective

## Develop a Predictive Model:

To create a reliable predictive model that could accurately forecast energy usage during July, the peak month of energy consumption. This model would help the company anticipate when and where high energy usage is likely to occur and adjust their energy supply accordingly. Additionally, it would identify times when the grid is at risk of being overloaded, allowing for preventive measures to be implemented.

## Understand Key Drivers of Energy Usage:

To analyze various data points related to energy consumption to identify the key drivers. These drivers could include but are not limited to, weather conditions (like temperature and humidity), household characteristics (such as house size, insulation quality, and the type and efficiency of appliances used), and behavioral patterns of residents. Understanding these factors would allow the company to tailor specific conservation strategies targeted at reducing consumption during critical times.

# Data Sources and Preparation

Our analysis utilized a comprehensive dataset encompassing:

## Static House Data:

Information on over 5,000 single-family houses, including attributes like building ID, square footage, and number of rooms.

## Energy Usage Data:

Hourly energy consumption for each of these houses, providing granular insights into usage patterns across different energy sources like AC and appliances.

## Weather Data:

Hourly weather information for each county in the service area, including temperature and humidity, which potentially impact energy usage.

These datasets were meticulously prepared through processes that included cleaning (handling missing data and outliers) and integration to align the various data sources into a cohesive structure conducive to analysis.

# Methodology

1. Data Collection:

The data collection phase of the project involved gathering diverse datasets from multiple sources, each contributing essential insights into the factors influencing energy consumption. The data collected can be broadly categorized into three main types: static house data, weather data, and energy usage data. Each dataset provided unique information critical for building an accurate predictive model and understanding the key drivers of energy usage. Static-House Data.

### Static-House Data

The static house data encompasses the structural and permanent characteristics of homes within the service area of the energy company. This dataset includes a variety of attributes that are fixed and do not change over time, making them static. These attributes are crucial as they provide a baseline understanding of the physical factors that may influence energy consumption patterns. Attributes included in the static house data:

Variables included in the Static-House data:

**Building ID (bldg\_id):** A unique identifier for each building, which helps in tracking and analyzing data specific to each structure.

**Square Footage (in.sqft):** The total area of the building, which is a significant factor as larger homes typically require more energy for heating and cooling.

**Number of Bedrooms (in.bedrooms):** Indicative of the potential occupancy and usage patterns within the home.

**Number of Stories (in.geometry\_stories):** This affects the energy usage due to factors like air circulation and space heating/cooling needs.

**HVAC Cooling Type (in.hvac\_cooling\_type) and Heating Type (in.hvac\_heating\_type):** The type of systems installed can significantly affect energy efficiency.

**HVAC Efficiency (in.hvac\_heating\_efficiency, in.hvac\_cooling\_efficiency):** Higher efficiency units consume less energy to achieve the same level of temperature control as less efficient units.

**Presence of Ducts (in.hvac\_has\_ducts):** Influences the efficiency of heating and cooling systems.

**County (in.county):** Geographic location can influence energy needs based on regional climate conditions.

**Number of Occupants (in.occupants):** More occupants generally lead to higher energy usage.

### Weather Data

Weather conditions play a crucial role in determining the energy consumption patterns, especially for heating and cooling. The weather dataset collected for this project includes hourly updates on various weather parameters across the counties served by the energy company.

Variables included in the weather data:

**Time (time):** Timestamps for each data entry, providing hourly updates.

**Dry Bulb Temperature [°C]:** The ambient air temperature, a critical factor in the energy consumption for HVAC systems.

**Relative Humidity [%]:** High humidity can increase the perceived temperature, affecting cooling systems' efficiency.

**Wind Speed [m/s] and Wind Direction [Deg]:** Wind conditions can influence the heating requirements and, in some cases, cooling patterns depending on the building's exposure.

### Energy Usage Data

The energy usage data provides detailed insights into the actual consumption of energy within each home. This dataset is vital for correlating the static and environmental factors with actual energy use outcomes.

Variables included in the energy dataset:

**Electricity, Fuel, Natural Gas, Propane:** These variables represent different energy sources used within the homes.

**Heating and Cooling:** Specific data on energy used for heating and cooling, which are typically the most significant energy consumption drivers in residential settings.

**Total Energy:** The aggregate energy usage, providing a comprehensive view of a household's total energy consumption.

## Data Analysis (EDA):

Conducting initial analyses to explore trends, patterns, and correlations within the data. This step helped in understanding the baseline energy usage behaviors and the impact of various factors on energy consumption.

### Distribution of Square Footage

A graph of a number of squares

Description automatically generated

This bar graph, titled "Distribution of Square Footage," illustrates the distribution of house sizes by square footage in a selected county. The x-axis indicates the square footage, segmented into ranges, while the y-axis shows the count of houses in each range. The graph reveals a peak in the 2000 square foot range, indicating this is the most common size of houses in the area. Fewer houses are found as the square footage increases, with a notably smaller count in the largest size category shown (around 8000 square feet).

This visualization helps us understand the typical house sizes in the county, showing that mid-sized homes are predominant. Such data is essential for analyzing market trends and planning in the housing sector, providing insights into what types of homes are most commonly available or preferred by residents in the area.

### Square Footage VS. Bedrooms by Number of stories

A graph showing a plot of a plot

Description automatically generated with medium confidence

This scatter plot titled "Square Footage vs. Bedrooms by Number of Stories" illustrates the relationship between the size of a house in square feet and the number of bedrooms, categorized by the number of stories in each house. The x-axis represents the square footage of houses, while the y-axis indicates the number of bedrooms. The color coding represents different numbers of stories, with specific colors corresponding to 1, 1.25, 1.5, 1.75, and 2-story houses.

### Wind Rose

A diagram of a circular object

Description automatically generated with medium confidence

This graph is a Wind Rose diagram, a specialized chart used to visualize the frequency and direction of wind over a given period. It displays how often winds blow from particular directions and the typical wind speeds associated with those directions. Each colored segment represents a different wind speed range, categorized from "0 to 2" meters per second up to "12.4" meters per second. The percentages on the outer rim indicate the frequency of wind from each direction.

From this diagram, we can deduce several insights:

- Wind is predominantly calm, as noted by the "calm = 11.5%" notation at the center.

- The mean wind speed is relatively low at 3.073 m/s.

- Wind direction varies, with no single direction overwhelmingly dominant, but there appears to be a slight preference for certain directions over others based on the distribution of the bars.

### Distribution of HVAC Heating type

A graph showing the number of hvac heating types

Description automatically generated

This bar graph, titled "Distribution of HVAC Heating Types," provides insights into the prevalence and energy consumption of different heating systems within a dataset. It categorizes heating into "Ducted Heat Pump," "Ducted Heating," "Non-Ducted Heating," and "None" for properties without heating. The graph clearly shows that "Ducted Heating" is the most common type, suggesting it might be the most favored or standard option. Although the graph does not explicitly detail energy consumption, understanding the distribution can hint at which systems might consume more energy based on their prevalence. "Ducted Heat Pump" is also notably common, while "Non-Ducted Heating" appears less frequently, and very few properties have no heating at all. The y-axis indicates the count of instances, and the x-axis lists the types of heating systems, providing a visual representation of heating trends in the dataset.

### Distribution of HVAC Cooling types

A graph showing different types of cooling type

Description automatically generated

This bar graph titled "Distribution of HVAC Cooling Types" provides insights into the frequency and potential energy consumption of different cooling systems within a dataset. It shows "Central A/C," "Heat Pump," "None," and "Room A/C" as the categories. The graph highlights "Central A/C" as the most common cooling system, suggesting it is likely the preferred choice in the dataset. "Heat Pump" and "Room A/C" are less common, with "Room A/C" being the least prevalent. The category "None" indicates properties without any cooling system.

The y-axis represents the count of instances, and the x-axis lists the types of cooling systems. Although this graph does not specifically address energy consumption, the prevalence of systems like "Central A/C," which typically consume more energy, could imply higher overall energy use associated with cooling in this dataset. This visual representation allows for easy comparison and understanding of cooling trends and their possible implications on energy usage.

### Energy Consumption Over time

A graph showing the amount of time

Description automatically generated

This graph, titled "Energy Consumption Over Time," displays the variation in energy usage for heating (in red) and cooling (in blue) systems from April to August. The x-axis measures time, while the y-axis quantifies energy consumption in kilowatt-hours (kWh).

Key observations from the graph include:

- Cooling energy consumption shows significant spikes during the warmer months, especially from late May through August, reflecting increased use of air conditioning systems as temperatures rise.

- Heating energy usage remains relatively low and consistent throughout the observed period, suggesting that there is little need for heating during these months.

- The overall energy consumption for cooling far exceeds that for heating, indicating a higher demand for cooling systems during the summer season.

This graph provides valuable insights into seasonal energy demands, highlighting the impact of climate and seasonal changes on energy usage patterns. It can inform resource planning and energy efficiency strategies for residential and commercial properties to manage peak energy loads effectively.

### Effect of Temperature on heating and cooling:

A graph showing the difference between heat and cooling

Description automatically generated

This graph, titled "Effect of Temperature on Heating and Cooling," plots the relationship between ambient temperature (in degrees Celsius on the x-axis) and energy consumption for heating (red line) and cooling (blue line) in kilowatt-hours (kWh on the y-axis).

From the graph, the following insights can be drawn:

- Cooling energy consumption (blue line) increases linearly with rising temperatures, indicating that more energy is used for cooling as it gets warmer.

- Heating energy consumption (red line) remains nearly constant and very low across the temperature range, suggesting minimal to no heating is required as temperatures increase.

- The significant difference in trends between the two shows that cooling needs are highly sensitive to temperature changes, whereas heating needs are not significantly affected within this temperature range.

This visualization helps illustrate how energy needs for cooling escalate with increasing temperatures, which is valuable for planning energy management and efficiency strategies, especially in regions with varying seasonal temperatures. It provides a clear demonstration of how ambient temperature impacts energy demands differently for heating versus cooling systems.

## Predictive Modeling:

In this project, we aimed to predict energy consumption specifically for heating and cooling during the month of July, a period marked by a 5-degree increase in temperatures. We focused on modeling the impact of this temperature rise on energy usage, employing two distinct predictive models: Linear Regression and XGBoost. Our models utilized dependent variables 'Heating energy' and 'Cooling energy' to gauge the energy consumption changes.

Data and Variables

The independent variables used in our models were chosen to effectively capture the factors influencing energy consumption for heating and cooling. These variables included:

- bldg\_id: Identifier for the building

- in.sqft: Square footage of the building

- in.bedrooms: Number of bedrooms

- in.geometry\_stories: Number of stories in the building

- in.hvac\_cooling\_type: Type of cooling system installed

- in.hvac\_heating\_type: Type of heating system used

- in.hvac\_heating\_efficiency: Efficiency rating of the heating system

- in.hvac\_cooling\_efficiency: Efficiency rating of the cooling system

- in.hvac\_has\_ducts: Presence of ducts in the HVAC system

- in.county: County where the building is located

- in.occupants: Number of occupants in the building

## Model Evaluation and Refinement:

Linear Regression Model Summary

The Linear Regression model served as our initial approach to understand how these variables correlate with energy usage. Despite its straightforward implementation and interpretability, the Linear Regression model achieved a limited accuracy of 64%. This performance suggests that while the model could identify linear associations between the features and energy consumption, it was insufficient for capturing more complex dependencies and interactions.

XGBoost Model Summary

To address the limitations observed with Linear Regression and better handle the non-linear relationships in the data, we employed the XGBoost model. Known for its powerful learning capabilities, XGBoost not only managed to incorporate the intricacies of the high-dimensional data but also improved the prediction accuracy to 68%. This model benefits from advanced features such as gradient boosting, tree pruning, and regularized learning, making it highly effective for this type of predictive task.

## Shiny App

The Shiny app developed for this project serves as a tool for visualizing energy consumption predictions alongside actual energy usage data. This application is specifically designed to provide insights into the energy consumption patterns for both heating and cooling during the month of July, which is the peak period of energy demand due to the increased need for cooling systems. The app includes interactive features allowing users to select specific date ranges to focus their analysis on particular intervals within the month.

The app presents two main types of visualizations:

1. Heating Energy Consumption: This graph displays the actual and predicted heating energy usage. Given the time of year (July), the heating energy remains consistently low with little to no spikes, as expected during the summer month.

A graph showing the energy consumption

Description automatically generated

1. Cooling Energy Consumption: This graph is more dynamic, illustrating both the actual and predicted cooling energy usage, which shows significant variability and higher energy consumption levels, reflecting the increased demand for cooling during hot weather.

A screenshot of a graph

Description automatically generated

# Key Findings

The results presented from the predictive modeling efforts using both linear regression and XGBoost models offer insights into the heating and cooling energy demands of residential areas serviced by the energy company. The models' effectiveness is primarily evaluated based on the R-squared metric, which provides a measure of how well the observed outcomes are replicated by the model

**Linear Regression Model Results**

**Heating Model**

R-squared: 0.6466161

The heating model under linear regression explains approximately 64.66% of the variance in heating energy usage. This suggests that while the model has a moderate predictive capability, there remains a significant portion of the variance that is unexplained by the model. This could be due to factors not captured in the model or inherent randomness in the energy usage data.

**Cooling Model**

R-squared: 0.6816466

The cooling model performs slightly better, explaining about 68.16% of the variance in cooling energy usage. This indicates a reasonably good fit, suggesting that the model captures a significant portion of the influences on cooling energy demands, possibly reflecting more consistent or predictable patterns influenced by variables such as temperature and humidity.

**XGBoost Model Results**

**Heating Model**

R-squared: 0.6559969

The XGBoost model for heating shows a slight improvement over the linear regression model, with an R-squared value of approximately 65.60%. This enhancement indicates the model's ability to handle complex non-linear relationships better than linear regression. The XGBoost model can pick up on interactions and patterns that the linear model might miss, providing a more nuanced understanding of the factors influencing heating energy consumption.

**Cooling Model**

R-squared: 0.6885249

For cooling, the XGBoost model explains about 68.85% of the variance, again showing an improvement over the linear regression model. This suggests that the XGBoost model, with its more sophisticated handling of data, is particularly effective in modeling cooling energy usage where complex interactions between variables are likely.

# Recommendations

1. Incentive Programs for Energy Conservation: Establishing a comprehensive incentive scheme is crucial for encouraging households to invest in energy-efficient technologies and make home improvements that conserve energy. Such programs should offer graduated benefits that cater to various income levels to ensure broad accessibility and effectiveness. By promoting energy efficiency, we can motivate a reduction in energy use across diverse economic backgrounds, yielding both financial savings and environmental advantages.
2. Adoption of Green Energy Technologies: Encouraging the use of renewable energy installations, such as solar panels, especially among wealthier households, plays a key role in diminishing overall energy demand. Financial incentives or tax rebates for investments in green technologies can enhance their appeal to financially well-off families. This strategy not only lessens dependency on conventional power sources but also cultivates a commitment to renewable energy within the community.
3. Educational Initiatives for Energy Conservation: Launching targeted educational initiatives aimed at affluent communities is crucial. These initiatives should highlight the environmental consequences of high energy use and offer actionable advice on reducing consumption. Educating residents about the benefits and practical steps towards energy conservation can empower them to make changes that have significant environmental impacts.
4. Progressive Energy Pricing Models: Implementing progressive pricing for energy consumption can be an effective deterrent against excessive use. By structuring prices so that they escalate with increased usage, households are motivated to implement energy-saving practices to keep their bills manageable. This pricing strategy not only curbs excessive energy use but also helps ensure a fairer distribution of energy costs.
5. Regulations and Building Codes for Energy Efficiency: Strengthening regulations and updating building codes to require energy-efficient standards in new constructions and major renovations can significantly reduce energy consumption. Mandating energy-saving materials and technologies in building design ensures energy efficiency from the ground up.
6. Community-Based Energy Programs: Developing community-based programs that leverage group purchasing for solar panels or energy-efficient appliances can reduce costs and barriers to entry for individuals. This collective approach can accelerate the adoption of energy-saving technologies in local communities, enhancing overall sustainability.
7. Smart Home Technology Integration: Promoting the integration of smart home technologies that automatically adjust energy usage based on real-time data can lead to significant reductions in household energy consumption. Incentives to adopt such technologies could be particularly appealing to tech-savvy, higher-income households.

By integrating these strategies, we can foster a reduction in energy consumption that aligns with broader environmental goals and promotes social equity. This holistic approach addresses the varying needs and capacities of different income groups, paving the way for a more energy-efficient and sustainable future.

# Conclusion

Our project has successfully developed predictive models that can significantly aid our client in managing energy demand effectively during critical periods. The models, built using linear regression and XGBoost techniques, provide robust forecasts of energy usage during the peak demand month of July. This capability is crucial for ensuring the stability of the energy grid during periods of high consumption due to increased cooling needs.

By identifying and understanding the key drivers of energy consumption through our analysis, such as HVAC efficiency, household size, and weather conditions, the energy company can implement strategic measures to mitigate the risk of grid overload. These measures not only help in managing immediate demand but also in planning for future infrastructure needs without excessive expenditure.

The implementation of our recommendations based on the predictive models could lead to significant enhancements in grid stability and energy efficiency. Such improvements not only contribute to reducing operational costs for the energy company but also support broader environmental sustainability efforts by minimizing unnecessary energy use and promoting energy conservation behaviors among consumers.

In conclusion, the predictive models and the insights derived from this project equip the energy company with the tools necessary to navigate the challenges of peak energy demand. This proactive approach ensures that the company can maintain a reliable energy supply, encourage sustainable energy usage, and move towards a more resilient and environmentally responsible energy management strategy.