INTRODUCTION

Link to R Shiny App: https://idsproject2023.shinyapps.io/IDS_Project_work/

Purpose

The purpose of this documentation is to perform comprehensive analysis to understand and predict energy consumption patterns for eSC. Through a blend of sophisticated data analytics and predictive modeling, this project aims to empower eSC with actionable insights that could lead to more efficient energy distribution, improved customer energy management, and the development of robust strategies to handle demand surges, particularly during peak usage times. The goal is to ensure eSC's infrastructure is resilient, responsive, and sustainable in the face of evolving energy demands.

Background

The backdrop of this study is the evolving landscape of energy needs against the broader canvas of technological advances and environmental imperatives. As South Carolina grapples with the dual challenge of fostering economic growth and reducing carbon footprints, eSC sits at the crux of this transformation. This project, therefore, is not just an academic exercise but a critical endeavor to distill insights from data that could shape the future of energy consumption and conservation in the region.

Methodology

Our methodology is a blend of data science techniques, starting with the amalgamation and cleansing of diverse datasets. We proceeded with an in-depth exploratory data analysis to identify patterns and anomalies. Predictive models were then constructed and validated to forecast hourly energy usage, incorporating potential scenarios like increased temperature effects. The results were synthesized into actionable insights through a user-friendly Shiny application, designed to facilitate interactive exploration of the data and predictions.

PROJECT DEFINITION

Problem Statement

The electric supply company eSC faces a critical challenge: managing energy distribution efficiently to meet fluctuating demands. With South Carolina's growing population and economic activities, the demand for electricity is surging, risking grid overload and energy waste. The challenge is compounded by the need to incorporate sustainable practices that mitigate environmental impact.

Objectives

The objectives of our project are multi-fold:

1. To conduct a granular analysis of historical energy consumption data to discern patterns and anomalies.
2. To build predictive models that can accurately forecast short-term energy demand, focusing on peak usage during the hot month of July.
3. To identify significant factors influencing energy consumption trends, including seasonal variations, economic activity, and consumer behavior.
4. To evaluate the potential impact of climate change on energy consumption by assessing scenarios with increased temperature forecasts.
5. To provide strategic recommendations that could lead to improvements in grid management and a reduction in peak load pressure.

Scope

The scope of this project is comprehensive and includes:

- **Data Integration:** Merging energy consumption data with external data sets such as weather conditions, housing characteristics, and demographic information to create a robust analytical framework.
- **Exploratory Data Analysis (EDA):** Utilizing statistical techniques to explore the data, uncover underlying structures, identify outliers, and hypothesize about potential relationships between variables.
- **Predictive Modeling:** Applying various statistical and machine learning techniques to build models that can predict energy usage, particularly in response to weather fluctuations.
- **Shiny Application Development:** Crafting an interactive application that visualizes the data and the predictive models' outputs, facilitating stakeholder understanding and strategic planning.
- **Future Forecasting:** Investigating how rising temperatures could alter consumption patterns and peak demand, providing eSC with a forward-looking perspective for infrastructure planning.
- **Documentation and Reporting:** Preparing detailed documentation that captures the methodology, analysis, results, and team contributions throughout the project lifecycle.

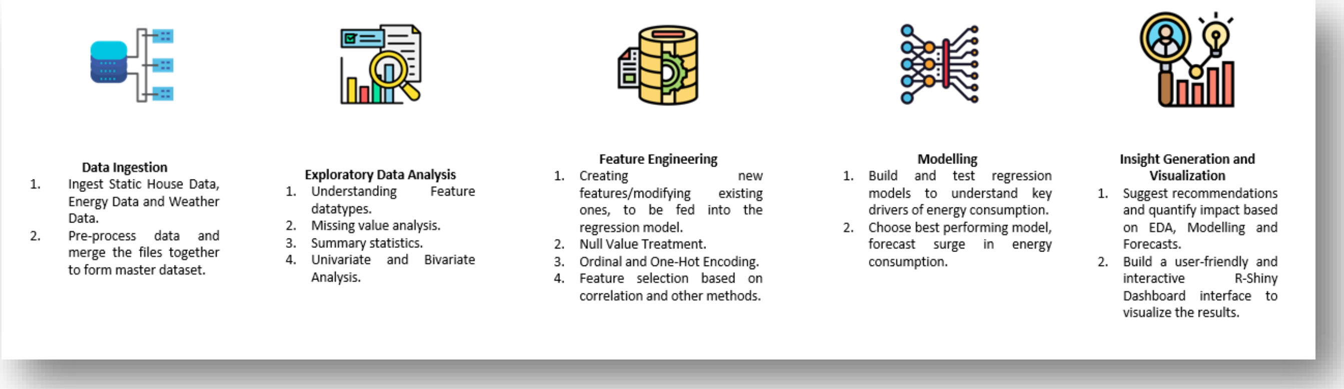
This project aims not only to present a snapshot of current energy usage but also to equip eSC with the tools and knowledge to anticipate and prepare for future energy needs.

Document Agenda

Addressed in the sections below are the answers to the following:

a) Determine the best approach to read and merge the data and determine what should be the output during this 'data preparation' phase.
b) Do exploratory analysis of the data – to gain some basic insight about the data
c) Build a model that predicts the energy usage, for a given hour, for the month of July. July was selected, as eSC thought July is typically the highest energy usage month. Hint: you will need to try several models and pick the best model.
d) Understand and be able to explain your model's accuracy.
e) Create a new weather dataset, with all July temperatures 5 degrees warmer
f) Use your best model to evaluate peak future energy demand (assuming no new customers)
a. Note: this must be model driven, not just increasing energy usage by a percentage
g) Show future peak energy demand in total (for an hour):
a. For different geographic regions
b. For other dimensions /attributes you think important
h) Create a shiny application so that your client can interact with the data
a. To better understand your model's energy prediction
b. To better understand the potential future energy needs
(and drivers of that future energy need)
i) Identify one potential approach to reduce peak energy demand
j) What would you suggest, how would you model the impact. How would you explain the impact.

RESEARCH AND DATA COLLECTION



Data Ingestion
1. Ingest Static House Data, Energy Data and Weather Data.
2. Pre-process data and merge the files together to form master dataset.

Exploratory Data Analysis
1. Understanding Feature datatypes.
2. Missing value analysis.
3. Summary statistics.
4. Univariate and Bivariate Analysis.

Feature Engineering
1. Creating new features/modifying existing ones, to be fed into the regression model.
2. Null Value Treatment.
3. Ordinal and One-Hot Encoding.
4. Feature selection based on correlation and other methods.

Modelling
1. Build and test regression models to understand key drivers of energy consumption.
2. Choose best performing model, forecast surge in energy consumption.

Insight Generation and Visualization
1. Suggest recommendations and quantify impact based on EDA, Modelling and Forecasts.
2. Build a user-friendly and interactive R-Shiny Dashboard interface to visualize the results.

Data Sources
The project taps into multiple data sources to gain a comprehensive view of energy consumption patterns. Primary data comes from eSC's historical records, detailing electricity usage across different consumer segments and time periods. Supplementary data includes weather information from meteorological services, capturing temperature, humidity, and precipitation, crucial for understanding weather-related energy consumption trends. Additionally, demographic and socio-economic data from government databases provide

insights into consumption patterns across different population segments. This multi-source approach ensures a holistic view of factors affecting energy consumption.

Data Processing Logic

The process of merging three datasets involved a logical sequence to ensure comprehensive integration. The first dataset, consisting of static house data, was retained at the building level without any alterations. The second dataset, containing energy data, was initially at an hourly level with multiple energy columns. To consolidate this information, all columns were summed to create a single energy column. The dataset was then filtered for the month of July, and the total energy for each building was computed for the entire month. Subsequently, this monthly energy data was divided into three distinct time periods of 8-hour intervals—morning, afternoon-evening, and night.

The third dataset, comprising weather data, was available at a county-hour level. Similar to the energy data, it was filtered for July, and the attributes for each county were averaged across the three designated time periods of 8 hours each—morning, afternoon-evening, and night.

In the final stages of integration, the static house dataset was joined with the energy dataset at the Building ID level. Further expansion of the dataset occurred with the inclusion of weather data, which was joined based on the county and time of day. This meticulous approach ensured a comprehensive and cohesive merging of the datasets, facilitating a more holistic analysis of the integrated information.

Data Cleaning

The data cleaning process in our project is comprehensive and multi-staged, ensuring the integrity and usability of the data. It begins with loading necessary libraries like 'arrow' and 'tidyverse'.

1. **Data Preparation:** The raw dataset is read from a CSV file, focusing on essential columns needed for the analysis. This step includes selecting relevant columns and filtering the data to include only those necessary for the project.
2. **Data Pruning:** The process involves removing unnecessary columns, such as "in.units_represented," to streamline the dataset for analysis.
3. **Handling Categorical Variables:** String columns are identified, and for each, the mean of total energy consumption per group is plotted, along with the count per group. This visual analysis aids in understanding the distribution and influence of categorical variables on energy consumption.
4. **Null Value Management:** Null values are identified and treated appropriately. For some columns, nulls are replaced with the most frequent value, while columns with a high number of nulls are removed entirely.
5. **Data Type Conversions:** Certain columns, like 'in.occupants', are converted to numeric to facilitate analysis.
6. **One-Hot-Encoding:** Using the 'caret' library, categorical variables are transformed into a format suitable for modeling.
7. **Correlation Analysis:** Correlation between variables and total energy consumption is calculated, helping to identify the most influential factors in energy consumption.

This meticulous data cleaning process is crucial to ensure the accuracy of the subsequent analysis and the reliability of the project's findings.

EXPLORATORY DATA ANALYSIS

Understanding Feature Data Types

In the initial phase of our Exploratory Data Analysis, we focus on discerning the nature of data we are working with, categorized broadly into numerical, categorical, and time-based data types. This classification is pivotal for guiding our subsequent analytical strategies.

1. **Numerical Data:** Our dataset comprises various numerical features, such as 'dry_bulb_temperature_[°c]' and 'wind_speed_[m/s]', representing continuous data. These features are integral for statistical computations and developing predictive models. Our analysis involves summarizing these numerical variables using descriptive statistics to understand their central tendencies and dispersions.
2. **Categorical Data:** Features like 'in.ceiling_fan' and 'in.geometry_floor_area_bin' are examples of categorical data in our dataset. They represent discrete categories, each with its own set of values. In handling categorical data, we employ techniques like one-hot-encoding, as shown in our R code, to convert these categories into a format suitable for machine learning algorithms.
3. **Time-Based Data:** The 'time_of_day' feature, a quintessential time-based data element, is crucial for analyzing patterns and trends over time. Time-based data is often used in trend analysis and forecasting models to understand how variables change over specific intervals.

Each data type requires distinct treatment and analysis techniques. For numerical data, we use summary statistics and visual plots to understand distribution patterns. Categorical data are processed to make them compatible with our modeling techniques. Time-based data are analyzed to unearth patterns and trends that could be pivotal in predicting energy consumption.

This comprehensive understanding of our dataset's features lays a solid foundation for our subsequent data cleaning, feature engineering, and predictive modeling steps, ensuring that our approach is tailored to the unique characteristics of each data type.

Missing Value Analysis and Treatment

In our project, a crucial step is the analysis and treatment of missing values, which ensures the quality and reliability of our dataset. We first identify columns with missing data. Our approach is two-fold: for some variables, like 'upgrade.water_heater_efficiency' and 'upgrade.cooking_range', missing values are replaced with the most frequent value, ensuring data consistency. In cases where columns contain a high number of missing values, such as 'upgrade.insulation_roof' and several others, we opt for removal to maintain data integrity. This careful treatment of missing data is essential to prepare our dataset for robust and accurate analysis in subsequent stages.

Summary Statistics and Univariate Analysis

In our project's Summary Statistics and Univariate Analysis phase, we delve deeply into each variable to understand its characteristics. Using R, we compute summary statistics for continuous variables like 'total_energy_consumption', 'dry_bulb_temperature_[°c]', and 'wind_speed_[m/s]'. These statistics include mean, median, mode, range, and standard deviation, offering a comprehensive view of data distribution.

For categorical variables, frequency distributions are analyzed. For example, we examine the distribution of features like 'in.ceiling_fan' and 'in.geometry_floor_area_bin' to understand how often different categories occur. This analysis helps in identifying dominant categories and patterns within the dataset.

The visual data from scatter plots and histograms, along with a correlation matrix, enriches our Summary Statistics and Univariate Analysis section significantly. These graphs exhibit relationships between variables such as 'dry_bulb_temperature_[°c]', 'relative_humidity_[%]', and 'in.cooling_setpoint' with the 'total_energy_consumption'. The scatter plots show trends and potential correlations, while the histograms of total energy consumption reveal the distribution of energy usage across the dataset.

The correlation matrix provides a numerical representation of how closely related the variables are to each other and to the total energy consumption. This is crucial in understanding which factors are most influential in predicting energy usage patterns.

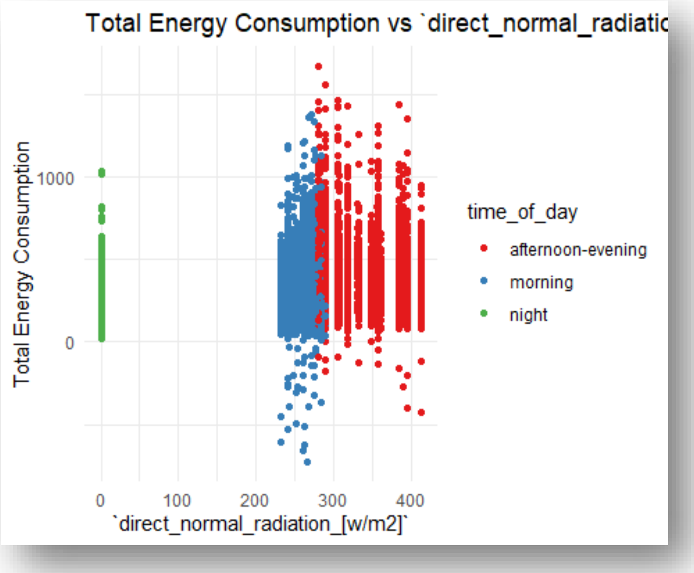
Histograms for continuous variables provide insights into the distribution shape, revealing any skewness or kurtosis. We observe data spread and central tendencies, which are crucial for identifying outliers or anomalies. The total energy consumption histogram turns out to be slightly Right Skew which allows us to focus on specific models while training the data.

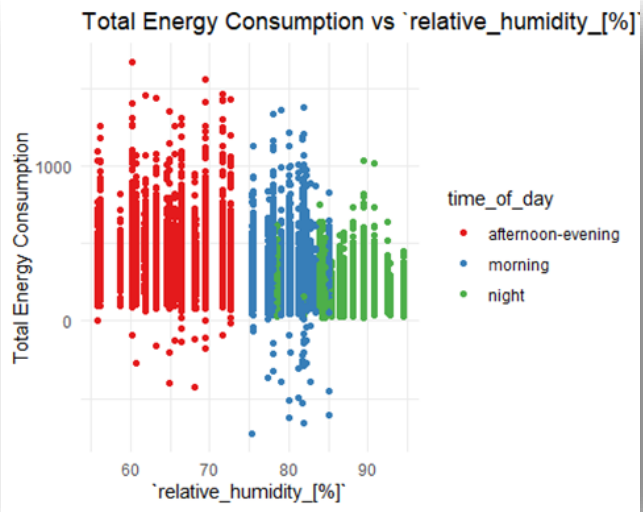
Furthermore, the bar plots depicting the average of the target variable by 'time_of_day' and by 'in.cooling_setpoint' offer an aggregated view of energy consumption trends across different times and thermostat settings. The variance in energy consumption during different times of the day is particularly telling, as it may indicate peak usage periods.

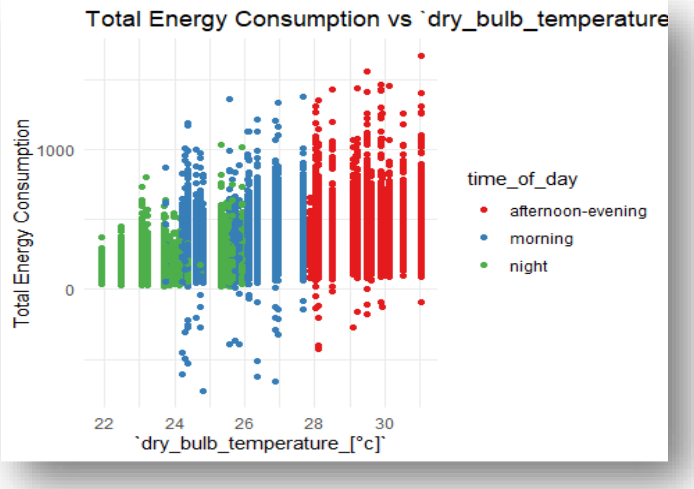
Lastly, the scatter plots color-coded by 'time_of_day' show how the relationship between temperature, humidity, wind speed, and energy consumption varies throughout the day. These visualizations are indispensable in our analysis as they help us understand the univariate properties of each variable, as well as their interactions, which are both fundamental to building accurate predictive models **(displayed below)**.

This comprehensive univariate analysis not only aids in understanding each variable in isolation but also prepares us for more complex multivariate analysis, ensuring a solid foundation for our predictive modeling and data interpretation.

The Detailed Analysis and Graphs for the same are present in the Exploratory Analysis Code File.







FEATURE ENGINEERING

Feature Engineering is a transformative process that turns raw data into features that better represent the underlying problem to the predictive models, resulting in improved model accuracy on unseen data.

1. **Creating and Modifying Features:** We have engineered features that capture more complexity than the original dataset. For example, 'time_of_day' was converted into a numerical variable to capture its cyclic nature, potentially enhancing the predictive model's ability to discern patterns over the course of a day. Additionally, we synthesized new features from existing ones, such as calculating the interaction terms between 'dry_bulb_temperature_[°c]' and 'relative_humidity_[%]' to understand their combined effect on energy consumption.

2. **Null Value Treatment:** Our approach to missing data is methodical and tailored to each feature's significance. Where appropriate, we have imputed missing values using the most frequent value or a central tendency measure like mean or median. This is evident in our treatment of 'upgrade.water_heater_efficiency', where missing entries were filled with the most frequent category, ensuring minimal bias in the dataset.

3. **Encoding Techniques:** We have converted categorical data into a numerical format using one-hot encoding. This technique transforms categorical variables into a form that could be provided to ML algorithms to do a better job in prediction. It was crucial for variables such as 'in.cooling_setpoint', which, although numerical in nature, represent discrete set points and are better treated as categorical.

4. **Feature Selection Strategies:** We leveraged correlation analysis and visual inspection of scatter plots and bar plots to determine the strength and nature of the relationship between variables. By doing so, we could identify and retain only those features that have the most significant impact on the target variable, 'total_energy_consumption'. This methodical approach to feature selection helps to streamline our model, reducing complexity and improving interpretability without compromising on performance.

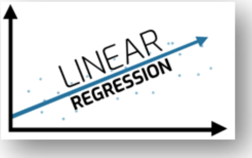
These steps in feature engineering allow us to refine our dataset into a more potent form for predictive modeling, setting the stage for more accurate and insightful outcomes from our machine learning algorithms.

PREDICTIVE MODELING

Model Development and Evaluation

Data Splitting: The dataset is judiciously split into training and testing sets, maintaining a balance that allows for adequate training of models while reserving a portion of data for unbiased evaluation.

- **Linear Regression**: We detail the implementation of Linear Regression, explicating its assumption of linearity and its limitations. The model's coefficients are interpreted to understand the impact of each predictor.
- **XGBoost**: This section elaborates on the use of XGBoost, highlighting its superiority in handling large datasets with complex, non-linear relationships. We explain the concept of gradient boosting and the specific hyperparameters tuned for this model.
- **CatBoost**: We describe the unique aspects of CatBoost, especially its adeptness in handling categorical data natively, which often presents a challenge in predictive modeling.



1. Accuracy : 75%
2. MAPE: 25
3. R-squared: 75%

1. Accuracy : 82%
2. MAPE: 18
3. R-squared: 82%

1. Accuracy : 82%
2. MAPE: 18
3. R-squared: 84%

Testing and Performance Metrics

- **Evaluation Approach**: Each model is rigorously tested using the test dataset. This section expands on the importance of using unseen data for model evaluation to prevent overfitting.

- **Performance Metrics**: We dive into each metric used, such as RMSE for its representation of the average error magnitude, MAPE for its percentage-based error measurement, and R-squared for its interpretation of variance explained by the model.

Model Selection and Performance Evaluation

- **Comparative Analysis**: A comprehensive comparison of the models based on the detailed performance metrics. This includes a discussion on the trade-offs made between model complexity and accuracy.
- **Shapley Value Analysis**: We provide an in-depth look at the use of Shapley values for model interpretability, explaining how this method decomposes a prediction into the sum of effects of each feature.

Conclusion

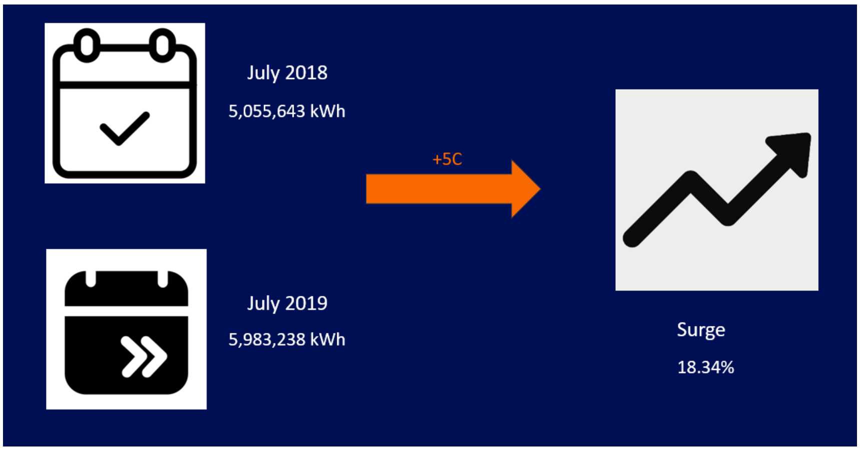
After a comprehensive evaluation of various models, the XGBoost model was selected as the optimal predictive model for our project. This decision was based on its exceptional performance metrics, notably achieving an R-squared value of 0.844847534851268. This high R-squared indicates that the XGBoost model explains approximately 84.48% of the variance in our target variable, showcasing its strong predictive power and accuracy. This selection underscores our commitment to leveraging advanced machine learning techniques to deliver precise and reliable predictive insights.

The Detailed Analysis and Graphs for the same are present in the Feature Engineering + Modelling Code File.

PREDICTIONS FOR JULY 2019

Detailed Analysis of Predictions

Upon increasing temperature by 5degrees C, and predicting next years temperature using our best performing model – XGBoost, we observed an 18.34% surge in next July.



July 2018
5,055,643 kWh

+5C

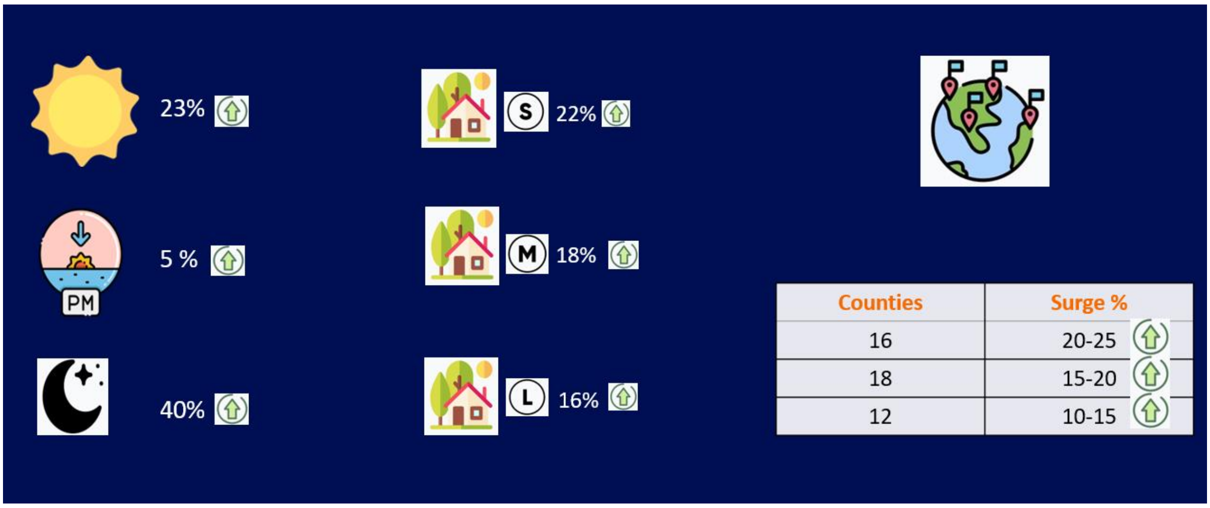
July 2019
5,983,238 kWh

Surge
18.34%

1.Broken down into geographies, we see that 16 counties will have a surge of 20-25%,18 counties will have a surge of 15-20% and 12 counties will have a surge of 10-15% in total energy consumption.

2.Broken down into time of day – mornings see a 23% surge, afternoon-evening a 5% surge and nights see a 40% surge in total energy consumption.

3.Broken down into house size – Small houses see a 22% surge, medium sized houses see a 18% surge and large houses see a 16% surge in total energy consumption.



23% ⬆

5 % ⬆

40% ⬆

S 22% ⬆

M 18% ⬆

L 16% ⬆

Counties	Surge %
16	20-25 ⬆
18	15-20 ⬆
12	10-15 ⬆

Using Shapley to generate and quantify insights

Shapley plots, derived from Shapley values, are a visualization technique used to interpret the contribution of individual features in a predictive model. Shapley values come from cooperative game theory and provide a way to fairly distribute the "value" of a coalition among its members. In the context of machine learning, the coalition consists of the features contributing to a prediction.

Here's how you can interpret Shapley plots:

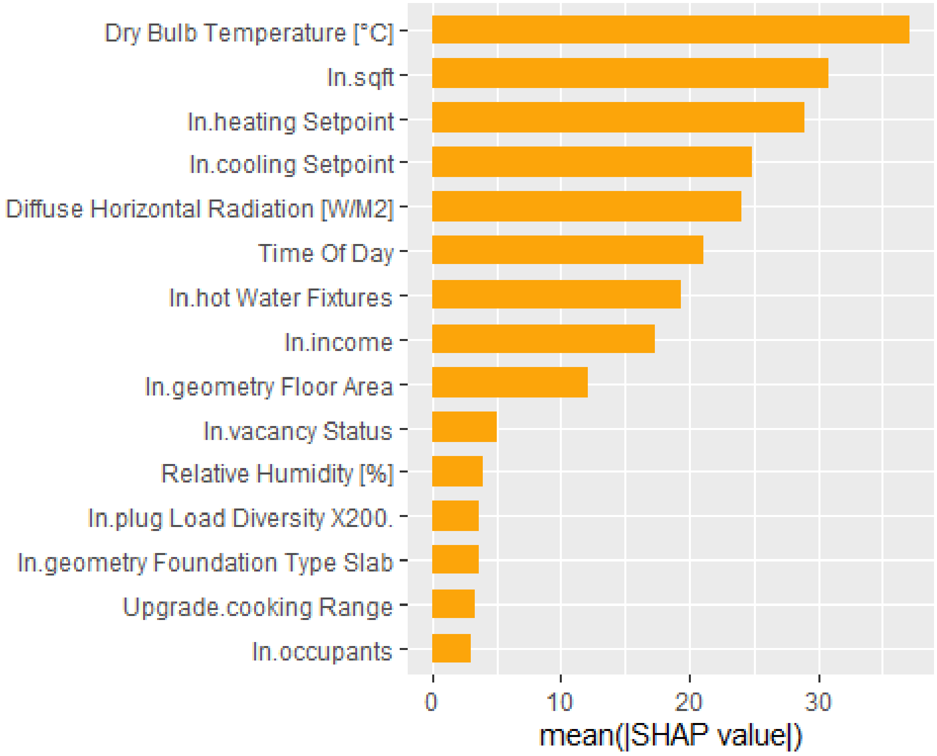
1. **Individual Feature Impact:**
   - Each point on the plot represents a single prediction.
   - The position of the point on the horizontal axis indicates the impact of the feature on the model's prediction for that specific instance.
   - Points to the right contribute positively to the prediction, while points to the left contribute negatively.
2. **Magnitude of Impact:**
   - The vertical distance from the reference point (usually the model's average prediction) represents the magnitude of the impact.
   - The longer the distance, the more the corresponding feature contributes to pushing the prediction away from the average.
3. **Color Coding:**
   - Some Shapley plots use color to represent the value of the feature for a specific instance.
   - Positive contributions might be colored differently from negative contributions, helping to distinguish the direction of impact.
4. **Model Output Interpretation:**
   - When examining Shapley plots for multiple predictions, you can get an overall sense of which features consistently contribute positively or negatively to the model's output.
5. **Interactions Between Features:**
   - Shapley values can also be used to explore interactions between features. For example, a high Shapley value for one feature might be mitigated by a high Shapley value for another.
6. **Summation to Prediction:**
   - The sum of the Shapley values for all features plus the average model output equals the model's prediction for a particular instance. This property allows you to understand how each feature contributes to the overall prediction.

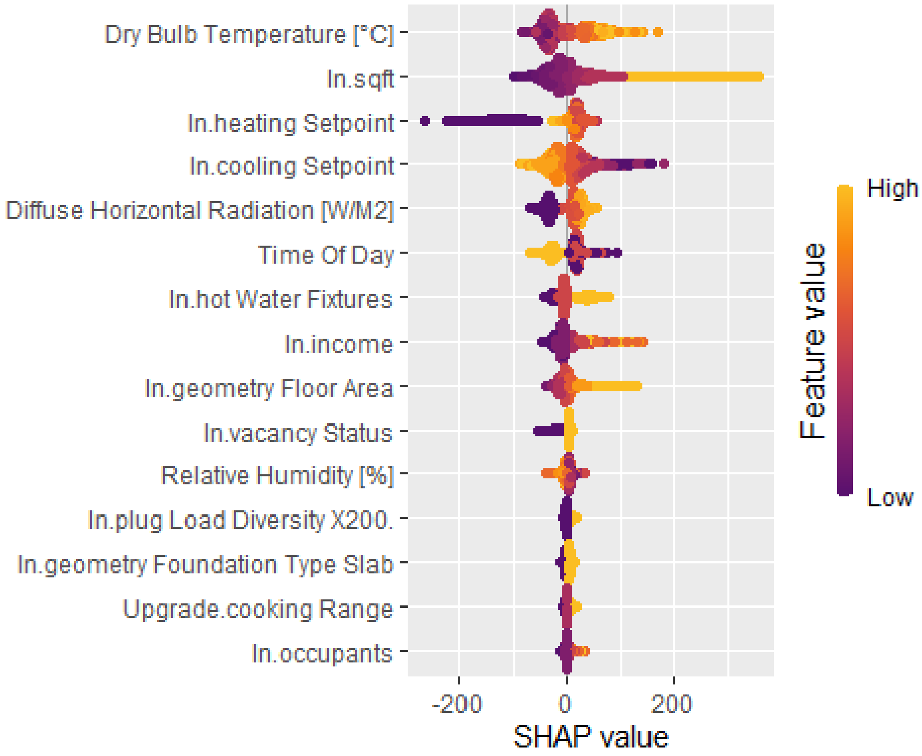
By analyzing Shapley plots, we gain insights into the importance and impact of individual features on the model's predictions. This interpretability can be valuable for understanding model behavior, identifying influential factors, and building trust in machine learning applications.

The Detailed Analysis and Graphs for the same are present in the Feature Engineering + Modelling Code File.

Feature Importance:

These graphs show a feature, in descending order of importance, and the impact they have on average energy consumption. For example, we see that dry bulb temperature in the dataset causes a 35 kWh deviation from the average total energy consumption as its value moves by 1 degree.





Partial Dependance plots

To study the marginal impact of the independent variable on the total energy consumption.

Partial Dependence Plot (PDP):
X-axis (Feature Values):

The horizontal axis of a PDP represents the values of the feature you are investigating.

Y-axis (Predicted Outcome):

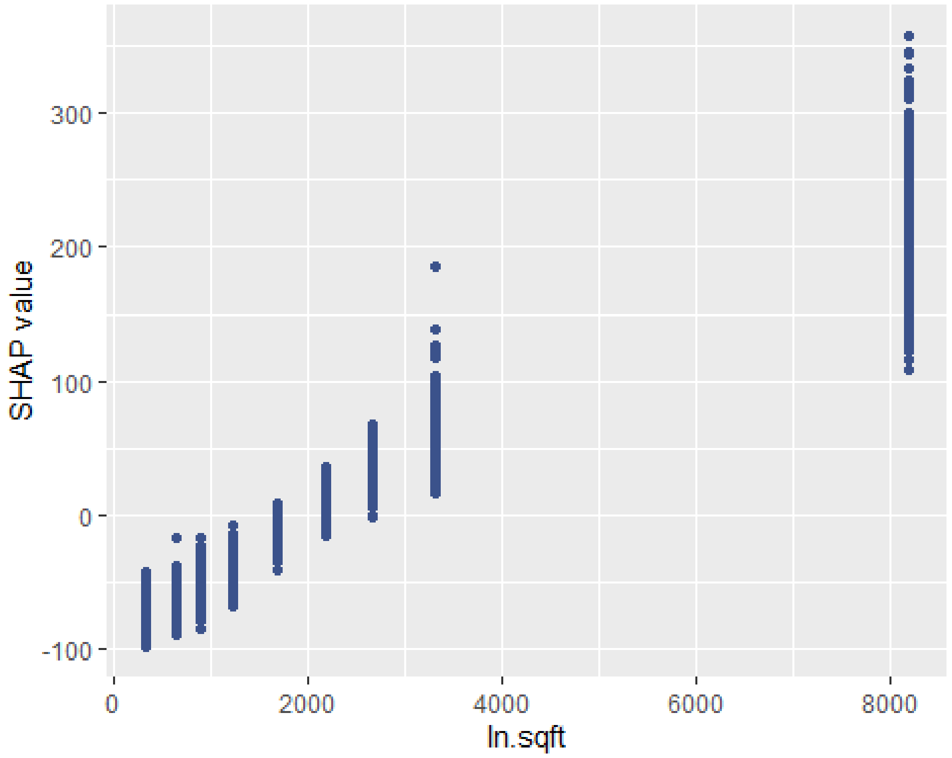
The vertical axis shows the predicted outcome (response variable) based on the feature values on the x-axis.

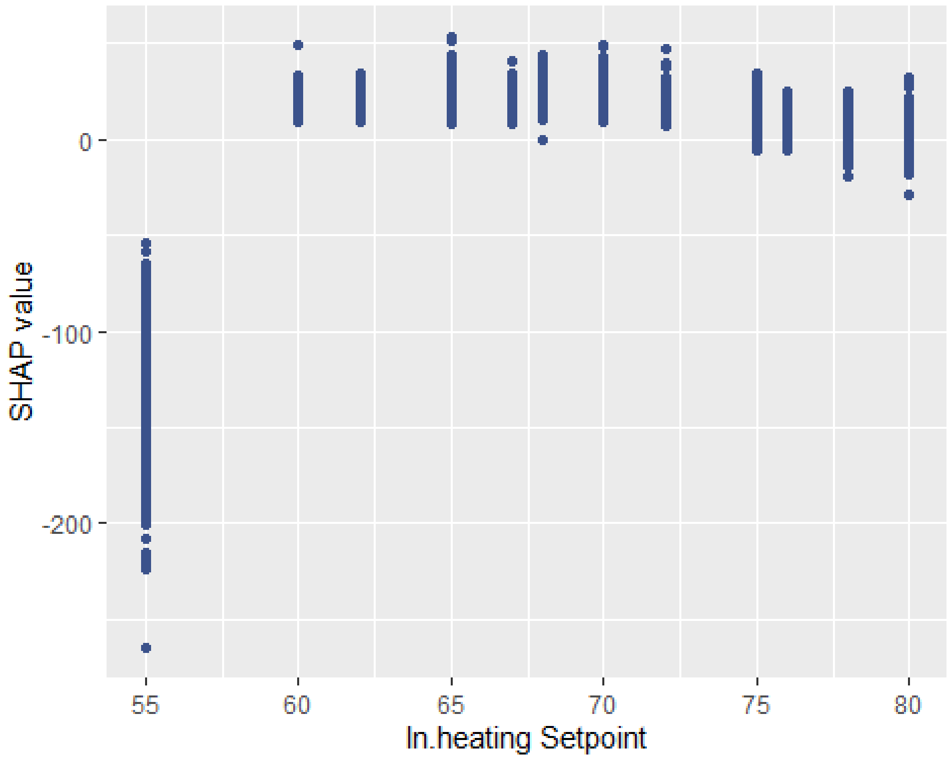
Main Curve:

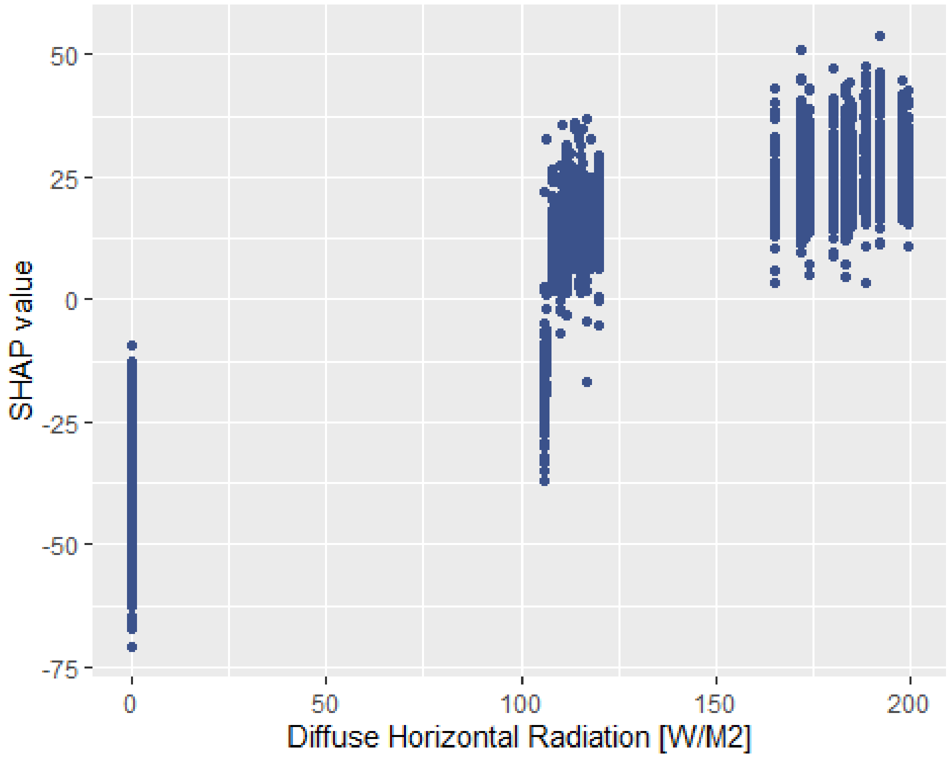
The main curve on the plot illustrates the average predicted outcome as the feature values vary. It helps to understand the general relationship between the chosen feature and the model's predictions while keeping other features constant.

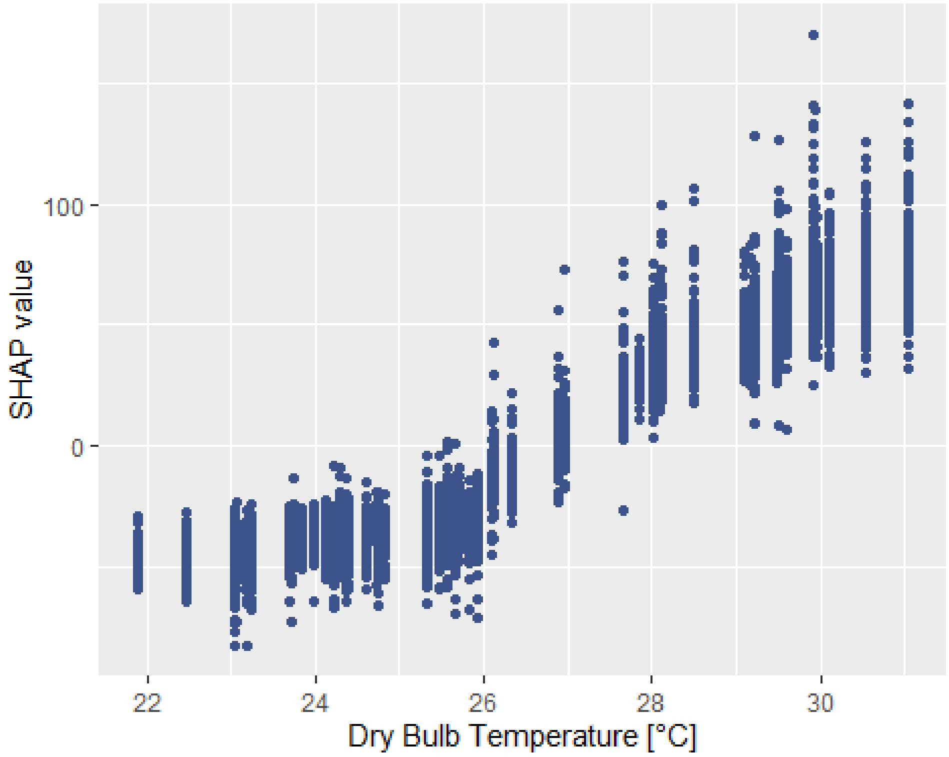
Shapley Values:
Shapley Values and Deviation from the Average:

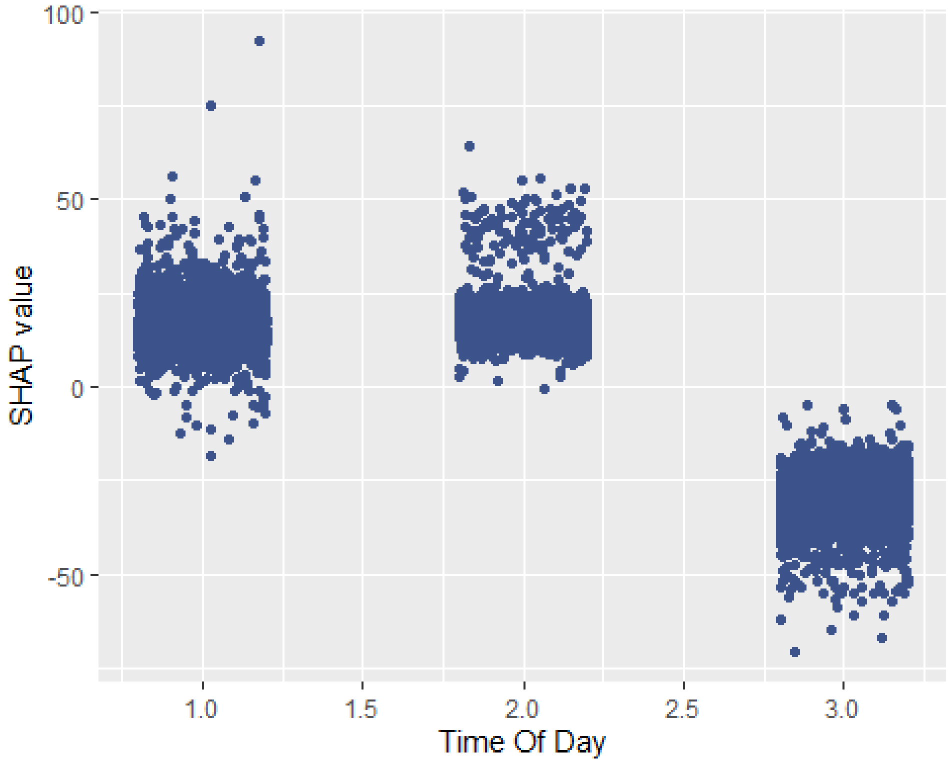
Positive Shapley values contribute to pushing the prediction above the average, while negative values contribute to pulling it below the average.

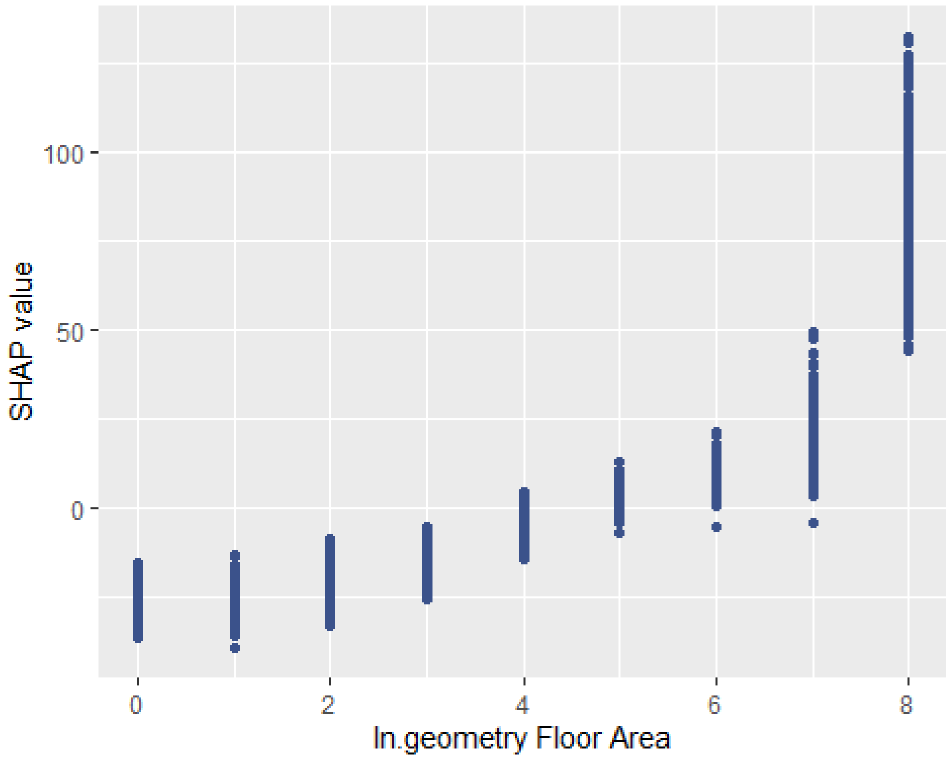

SHAP value
300
200
100
0
-100
0
2000
4000
6000
8000
In.sqft

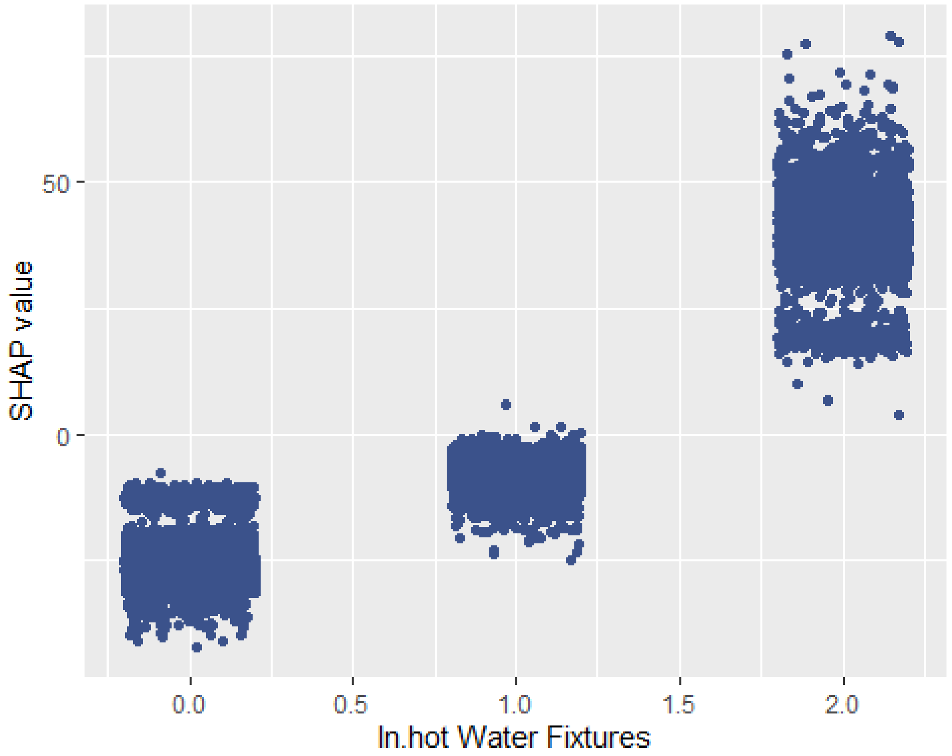

SHAP value
0
-100
-200
55
60
65
70
75
80
In.heating Setpoint

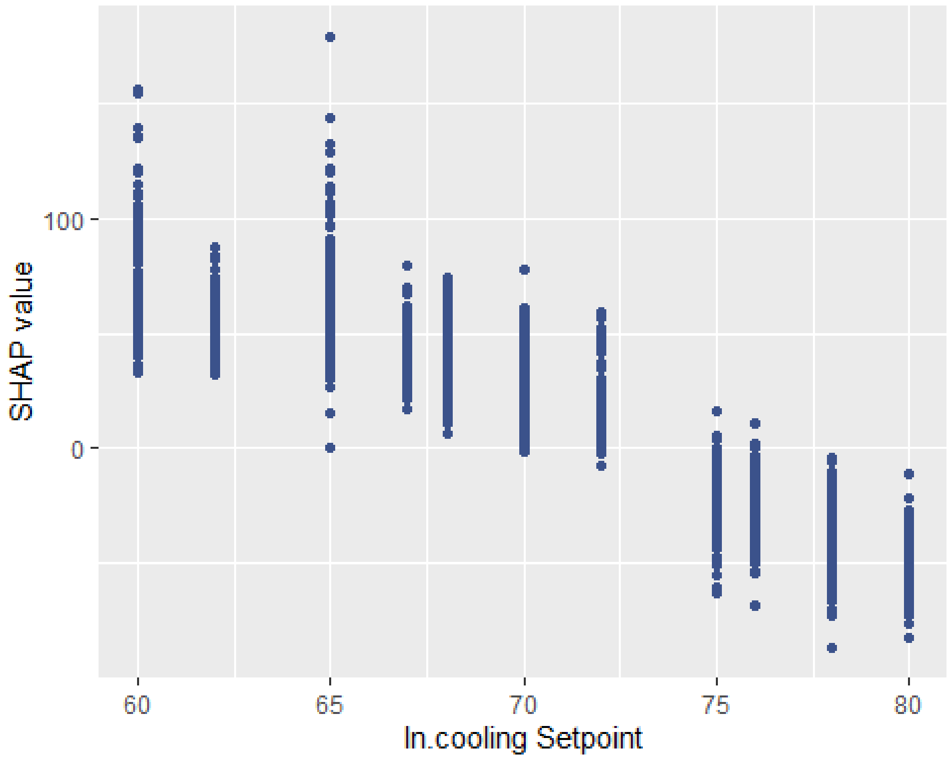

50
25
0
-25
-50
-75
SHAP value
0
50
100
150
200
Diffuse Horizontal Radiation [W/M2]

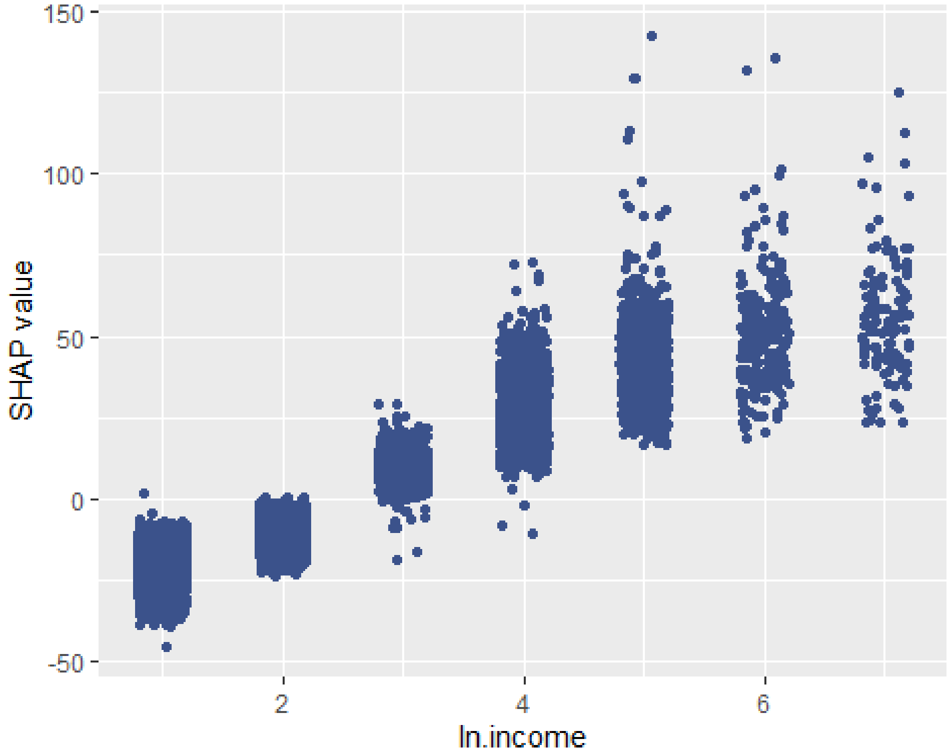

100
0
SHAP value
22
24
26
28
30
Dry Bulb Temperature [°C]

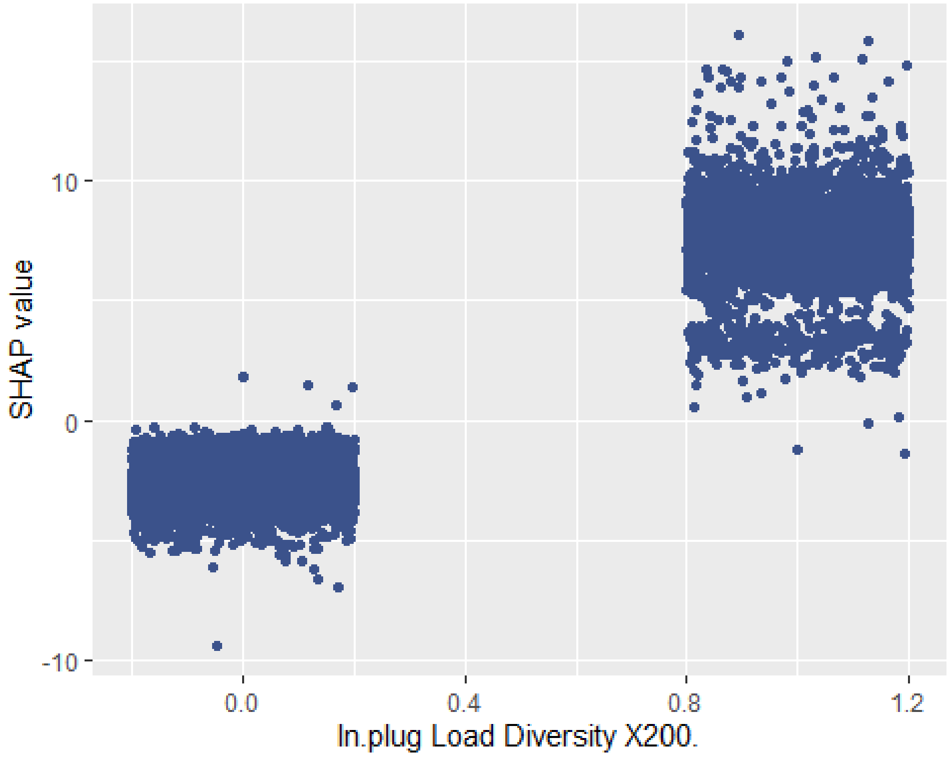

100
50
0
-50
SHAP value
1.0
1.5
2.0
2.5
3.0
Time Of Day

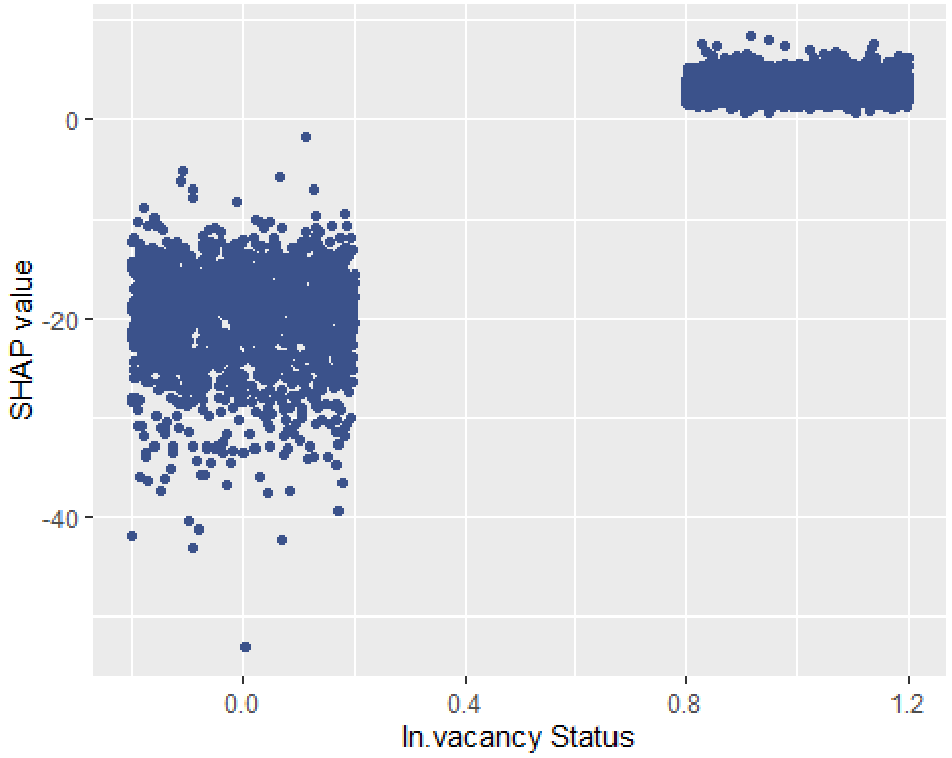

100
50
0
SHAP value
0
2
4
6
8
ln.geometry Floor Area

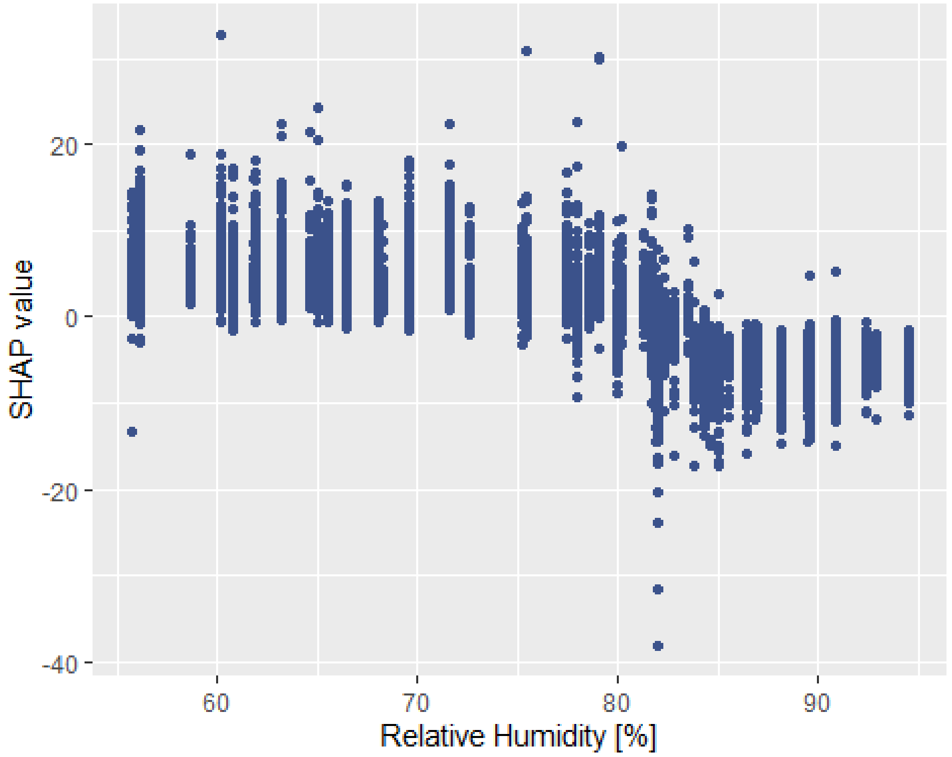

SHAP value
50
0
0.0
0.5
1.0
1.5
2.0
In.hot Water Fixtures

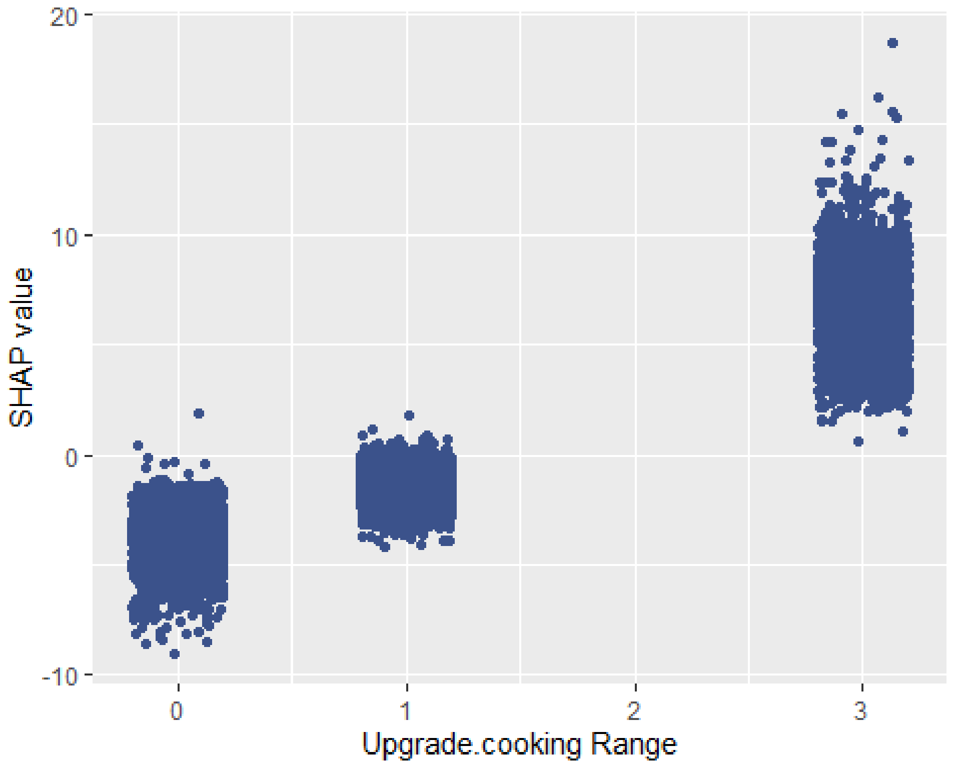

SHAP value
100
0
60
65
70
75
80
In.cooling Setpoint

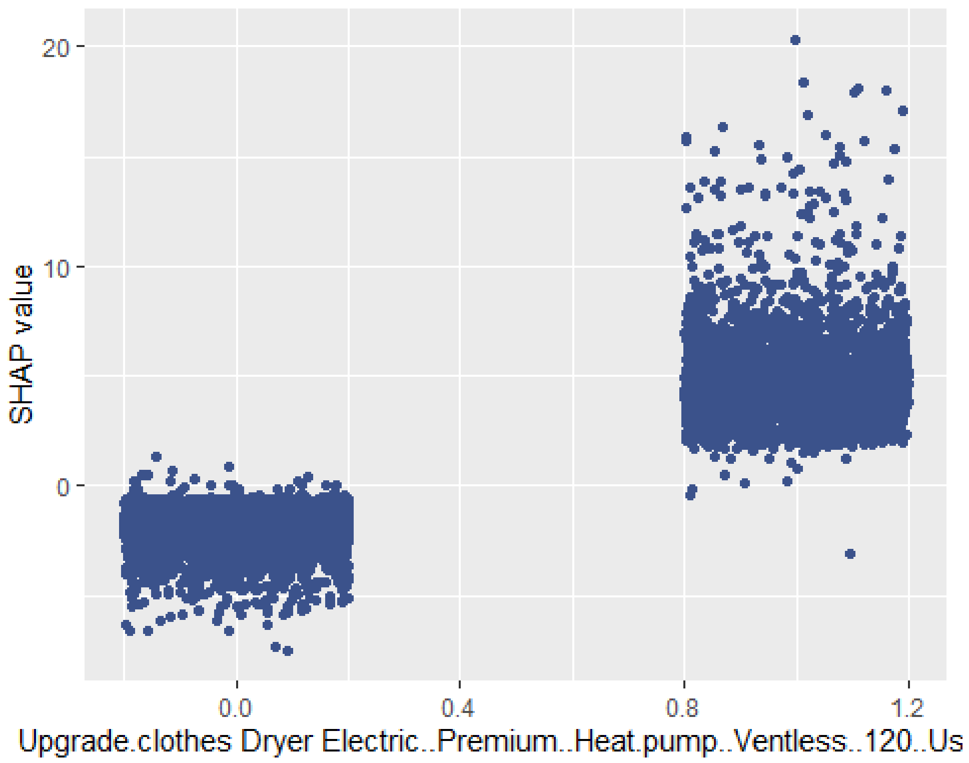

150
100
50
0
-50
SHAP value
2
4
6
ln.income

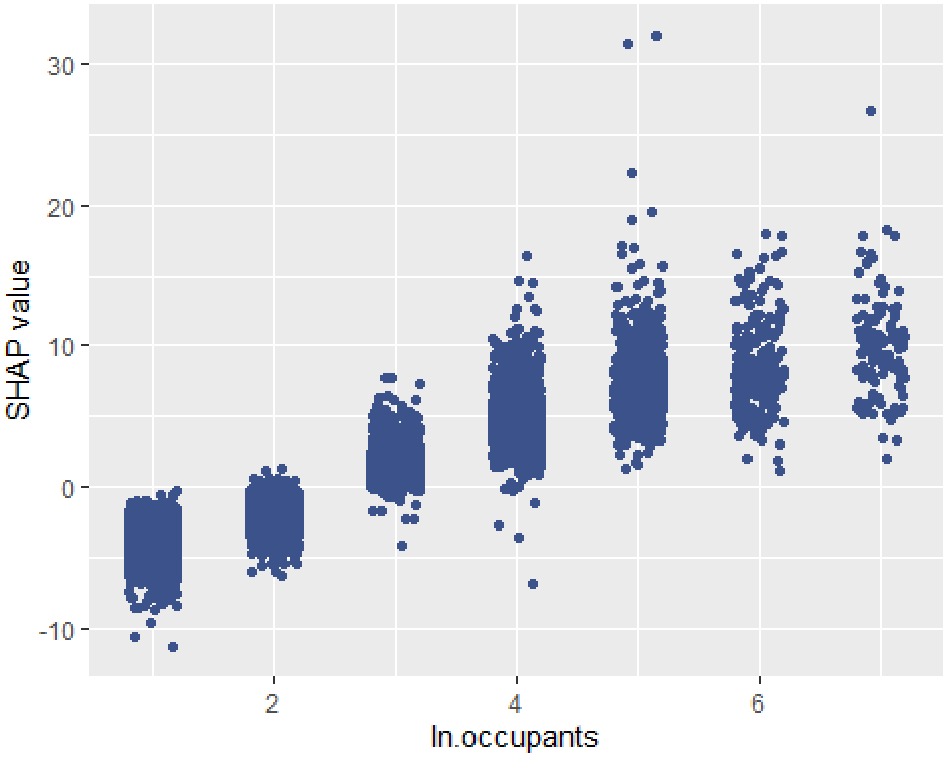

10
0
-10
SHAP value
0.0
0.4
0.8
1.2
ln.plug Load Diversity X200.


0
-20
-40
SHAP value
0.0
0.4
0.8
1.2
In.vacancy Status


20
0
-20
-40
SHAP value
60
70
80
90
Relative Humidity [%]


20
10
0
-10
SHAP value
0
1
2
3
Upgrade.cooking Range


20
10
0
SHAP value
0.0
0.4
0.8
1.2
Upgrade.clothes Dryer Electric..Premium..Heat.pump..Ventless..120..Us


SHAP value
30
20
10
0
-10
2
4
6
ln.occupants

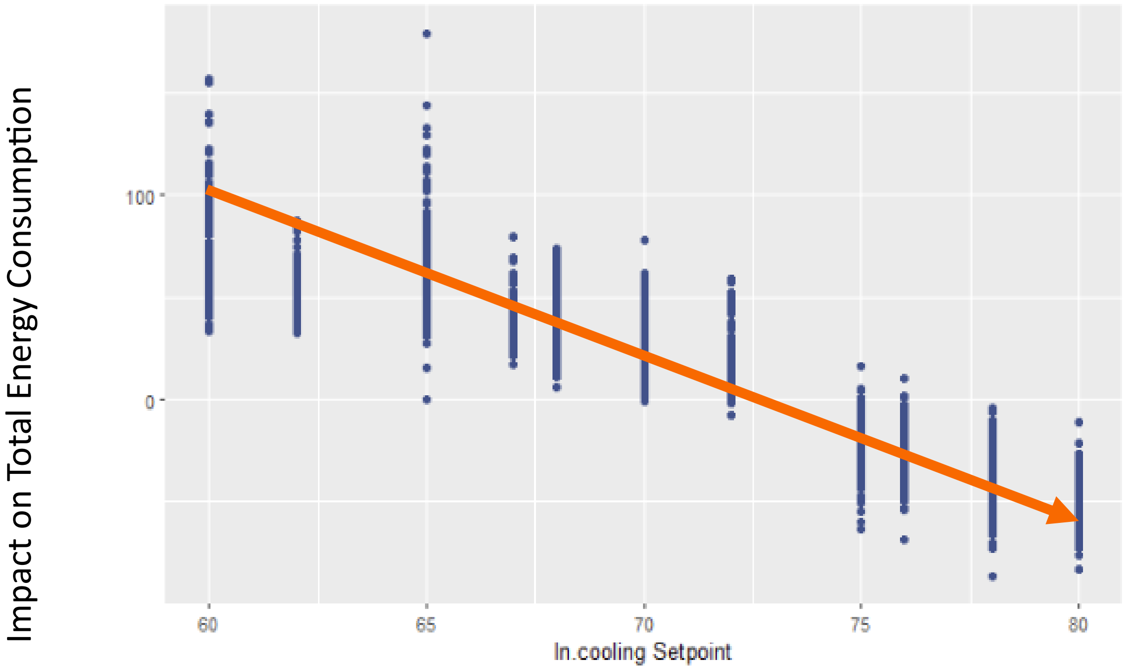
RECOMMENDATIONS AND STRATEGIC INSIGHTS

Recommendation 1

Launch campaigns to encourage adoption of Smart Appliances in buildings to **dynamically control cooling setpoint.**

Impact: **Impact has been calculated using Shapley Partial Dependence Plots.**

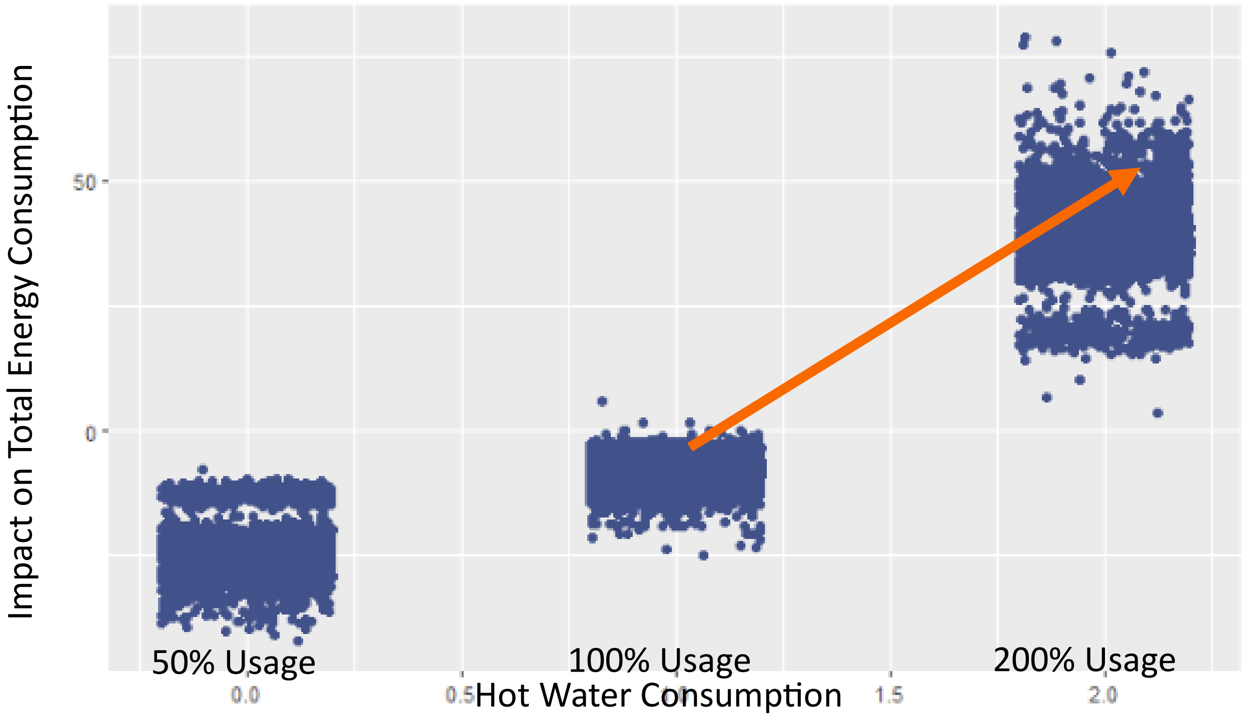
There is an average impact of **- 18.75 kWh per building per degree Fahrenheit** as the building thermostat temperature increases from **60F (15C) to 80CF (26C).**



Recommendation 2

Conduct state wide campaigns to reduce **Hot Water Fixture Costs**. Ensure fix leaks, install low-flow fixtures, insulate accessible hot water lines, and purchase an ENERGY STAR certified dishwasher and clothes washer.

Impact: **Impact has been calculated using Shapley Partial Dependence Plots.**
225 kWh of energy per building can be saved by reducing fixture usage from 200% to 100%.



VISUALIZATION

Development of R-Shiny Dashboard

Link to R Shiny App: https://idsproject2023.shinyapps.io/IDS_Project_work/