Introduction

Project Context and Motivation

In the face of mounting environmental challenges, particularly the intensification of global warming, energy providers must anticipate and adapt to rapidly changing patterns in energy demand. The southeastern United States, including South Carolina and parts of North Carolina, is especially vulnerable to extreme heat events during the summer months.

eSC, a regional electricity provider, has commissioned this initiative to address the anticipated surge in energy consumption, especially under hypothetical extreme weather scenarios. Specifically, the company is concerned about the possibility of significantly elevated temperatures during the month of July, traditionally the period of peak electricity usage. The risk is that a 5°C increase in average temperature could lead to energy demands that surpass current infrastructure capacity, potentially resulting in widespread blackouts.

Instead of investing in costly and environmentally impactful infrastructure expansions—such as building a new power plant—eSC is exploring data-driven alternatives. The goal is to deeply understand the factors driving peak energy usage and to forecast future demand under warmer conditions. This foresight would enable targeted energy-saving interventions, allowing the company to ensure uninterrupted service.

Objective of the Project

The objective of this project is to build a predictive framework that can:

- Accurately model hourly residential electricity usage during the month of July,
- Simulate how that usage would shift under a +5°C warming scenario,
- Identify peak energy usage periods and households most likely to contribute to these peaks,
- Explore regional differences and household characteristics that explain demand variability,
- Recommend actionable strategies for reducing or shifting usage to alleviate pressure on the grid.

Data-Driven Consultancy Role

In this project, our team acts as consultants to eSC, bringing a comprehensive data science approach to an applied energy systems challenge. The approach involves:

- Integrating different data sources including static housing attributes, hourly energy usage data, and weather information at the county level,
- Cleaning and merging huge datasets programmatically,
- Developing and comparing several predictive models (including linear regression, SVMs, Random Forests, and XGBoost),
- Performing predictive analysis through climate simulation,
- Designing a user-friendly Shiny application to democratize model interaction and visualization,
- Delivering a full technical report with business-friendly insights.

Business Questions

In alignment with eSCs strategic imperative to maintain grid reliability amidst a warming climate, this project is designed to address a focused set of data-driven business questions. These questions directly inform eSCs operational planning, infrastructure decisions, and customer engagement strategies for the upcoming summer and beyond. They represent the bridge between technical modeling work and tangible business outcomes.

Primary Business Question

How will residential energy usage in July change if temperatures rise by 5°C, and can eSC meet this increased demand without expanding its energy infrastructure?

This question is at the core of the project. It calls for a thorough forecast of hourly energy usage under a simulated climate change scenario. The analysis must produce both individual-level and system-wide (aggregated) demand projections, especially for peak hours, which are most critical to grid stability.

What are the key drivers of hourly energy usage in residential homes during July?

- Which household features (e.g., floor area, insulation, HVAC presence) most influence consumption?
- How do weather conditions (e.g., temperature, humidity) and time-of-day patterns contribute to usage variation?

This informs eSC where to focus efficiency programs—e.g., targeting large, poorly insulated homes with high cooling demand.

How accurately can energy usage be predicted at an hourly level for individual homes, given static house features and weather inputs?

- What is the best-performing model (e.g., Random Forest, XGBoost, SVM, Linear Regression)?
- How well does the model generalize to new houses or weather conditions?
- A high-accuracy model enables confident forecasting and scenario planning.

Under a +5°C warming scenario, when will the system experience its peak demand, and what will the magnitude of that peak be?

- On which day and hour is the peak likely to occur?
- How does this vary across counties or climate zones?

This guides operational readiness and contingency planning, such as scheduling demand-response events or peak-shaving programs.

What specific energy-saving strategies can be recommended to reduce peak demand, and what would be their projected impact?

This answers eSCs request for actionable insights—how to reduce demand instead of expanding capacity.

How can results be made accessible to non-technical stakeholders at eSC?

- What interactive visualizations, data drill-downs, and predictive tools should be included in a Shiny web application?
- How should model outputs be communicated to ensure interpretability and strategic alignment?

The Shiny app functions as a decision support tool for internal planning and executive communication.

Together, these business questions ensure that the project is not merely a technical exercise, but a strategic engagement aimed at transforming data into decisions. By answering them, eSC gains the insight needed to mitigate climate risks, maintain operational integrity, and deliver environmentally responsible service—all while avoiding capital-intensive expansions.

Data Acquisition and Preparation

The datasets provided by eSC consist of lots of different types of data—static housing attributes, time-series energy consumption, weather conditions, and an extensive metadata dictionary. This section outlines how these datasets were programmatically merged, cleaned, and validated to form a foundation for all subsequent modeling and simulation tasks.

Data Sources Overview

The following datasets were integrated:

Static House Data

A dataset with 5,000+ rows, each representing a residential property serviced by eSC. Attributes include unique building IDs (building_id), physical characteristics (e.g., in.geometry_floor_area, in.insulation_wall), location (in.county), and system-level indicators (e.g., presence of cooling equipment).

Energy Usage Data

A directory of 5,000+ parquet files, each named after a building_id. These files contain granular, hourly energy usage data for each home, with disaggregated source-level consumption (e.g., HVAC, dryers, appliances).

Weather Data

Hour-by-hour meteorological data for each county served by eSC, stored in ~50 csv files named by county code (e.g., G4500010.csv). Fields include temperature, humidity, dew point, and more.

Metadata Dictionary

A data_dictionary.csv file detailing the schema and description of approximately 270 attributes used across the datasets.

Data Access Methodology

Given the volume and format of the data, the following approach was adopted in R:

- Used the arrow package to read parquet files efficiently.
- Automated retrieval of individual house and weather files using scripted logic.
- Began the project using a single building (building_id = 102063) to validate the end-to-end pipeline, later scaling to batches of buildings for broader generalization.

- Weather data was matched to homes using the `in.county` field from the static dataset.

Data Merging Strategy

Merging the datasets required a consistent and modular approach:

- Static + Energy Data: Merged on building_id. Energy data was loaded for July only (744 hours). Timestamp formatting (POSIXct) was standardized to allow seamless joins.
- Static + Weather Data: Joined based on in.county, aligning each house with its corresponding county's weather records.
- Energy + Weather Data: Combined using both timestamp and county. Care was taken to ensure alignment at the hourly level—critical for time-series prediction.

An intermediate "working dataset" was created for each house, containing:

- Energy usage (target variable)
- Static house attributes
- Hourly weather metrics
- Temporal features (hour of day, day of week)

Data Cleaning and Transformation

The raw datasets required several steps before analysis:

- Missing Values
    - Static attributes with missing values were either replaced with median values (for numerical fields) or flagged with binary indicators.
    - Time gaps in energy or weather data were forward-filled or dropped if exceeding a threshold.
- Timestamp Normalization
    - Converted all timestamps to the same timezone.
    - Verified time coverage: each house's energy data had 744 hourly entries for July.
- Filtering and Scoping
    - Limited all analysis to July data only, to reflect peak usage month
    - Removed any buildings with incomplete energy records for July

Validation and Quality Checks

Several quality control steps were executed:

- Row Count Validation: Ensured that each merged dataset contained exactly 744 hourly rows per building
- Range Checks: Verified that temperature values fell within plausible July ranges (20–40°C) and checked that energy consumption values were non-negative and realistic

- Cross-Dataset Consistency: Validated join integrity and confirmed 1:1 mapping between time entries in energy and weather files

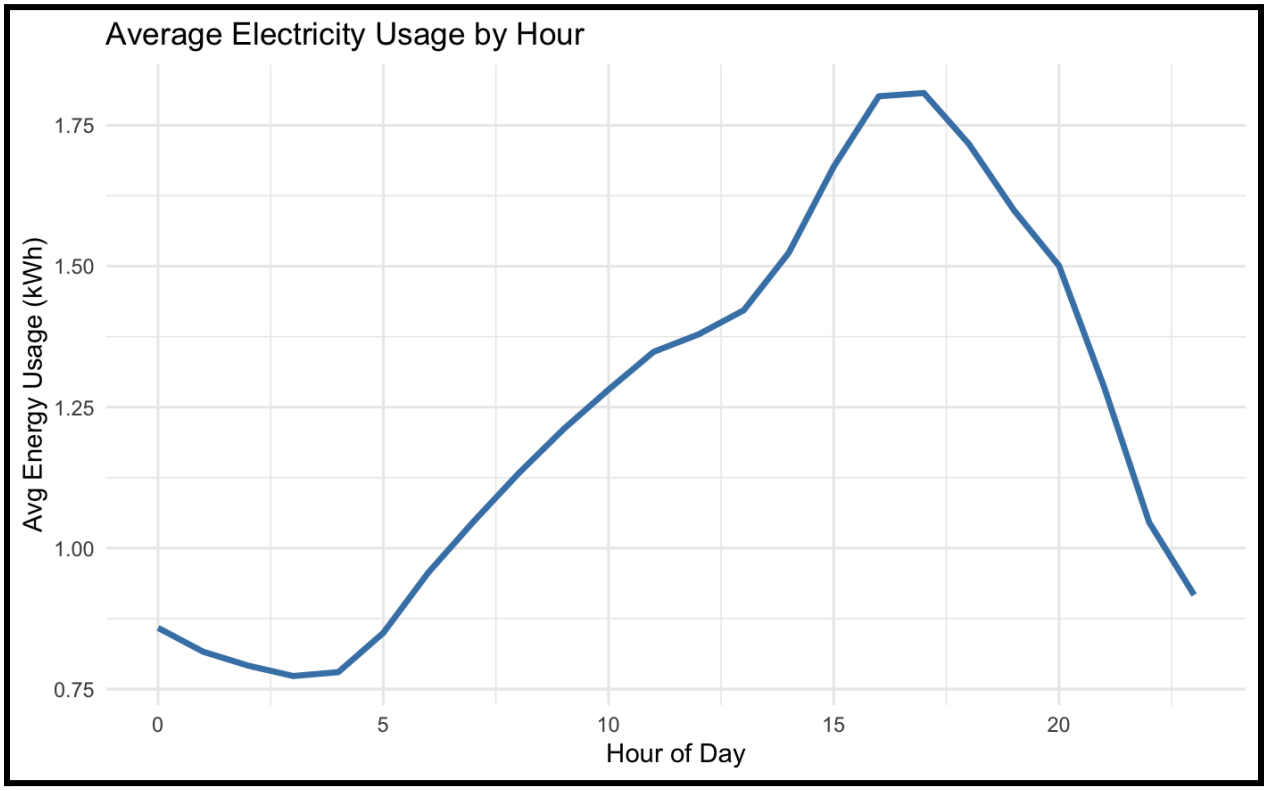
Exploratory Data Analysis (EDA)

The exploratory phase focused on identifying temporal, geographic, climatic, and structural patterns in July energy consumption.

Temporal Patterns: Hourly Energy Usage

- A clear pattern emerged, with energy usage rising sharply in the afternoon and peaking between 4 PM and 6 PM.
- Usage was lowest in the early morning hours, averaging around 0.75–0.85 kWh, and increased steadily through the day to above 1.8 kWh at peak.
- This pattern aligns closely with outdoor temperature trends and residential behaviors (such as air conditioning during home return hours).

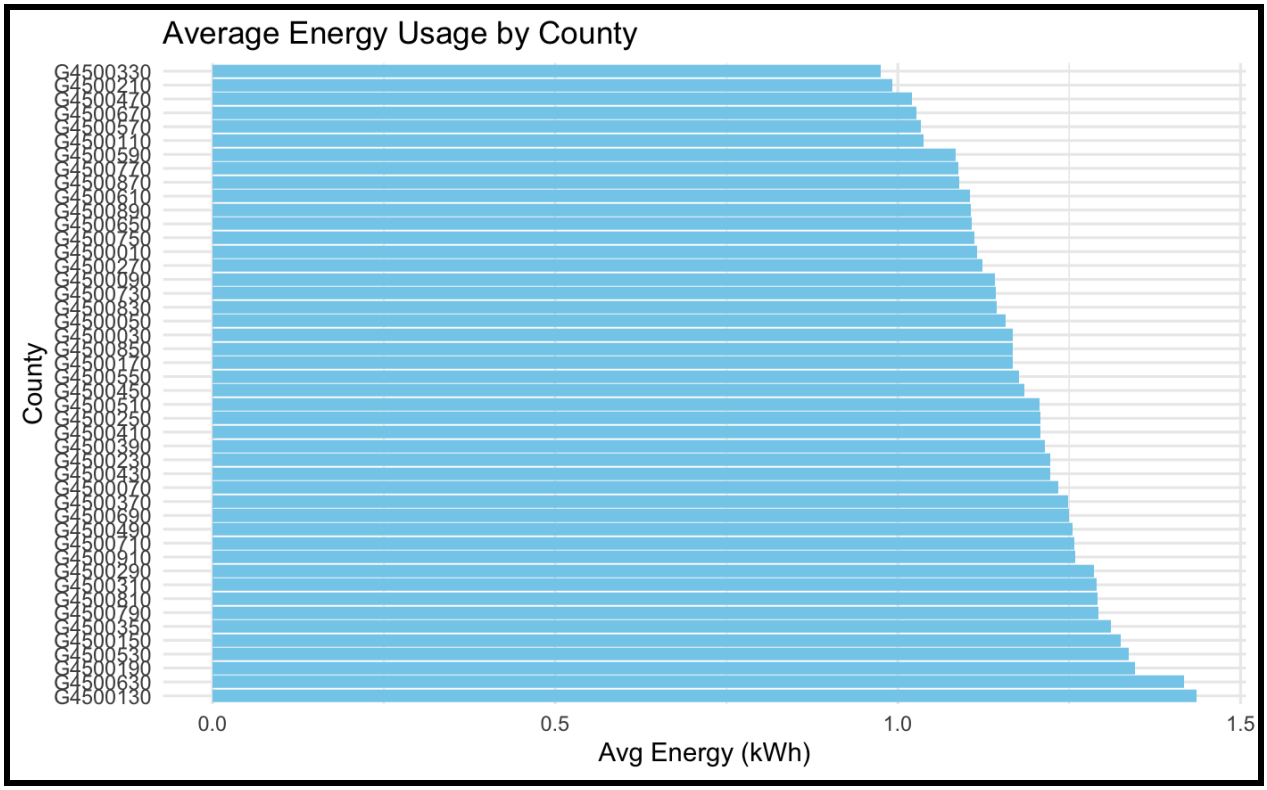
This plot confirms the late afternoon as the most critical period for energy planning, due to the convergence of peak temperature and residential load.



Geographic Patterns: County-Level Variation

- Average hourly energy usage varied significantly across counties, even after adjusting for floor area.
- Counties such as G4500130 and G4500630 showed the highest average usage, close to 1.5 kWh, while others remained below 0.75 kWh.
- These differences likely reflect climatic variations, building stock characteristics, and socioeconomic factors such as home size or appliance use.

This horizontal bar chart highlights counties that may need targeted demand-management strategies or infrastructure reinforcement.

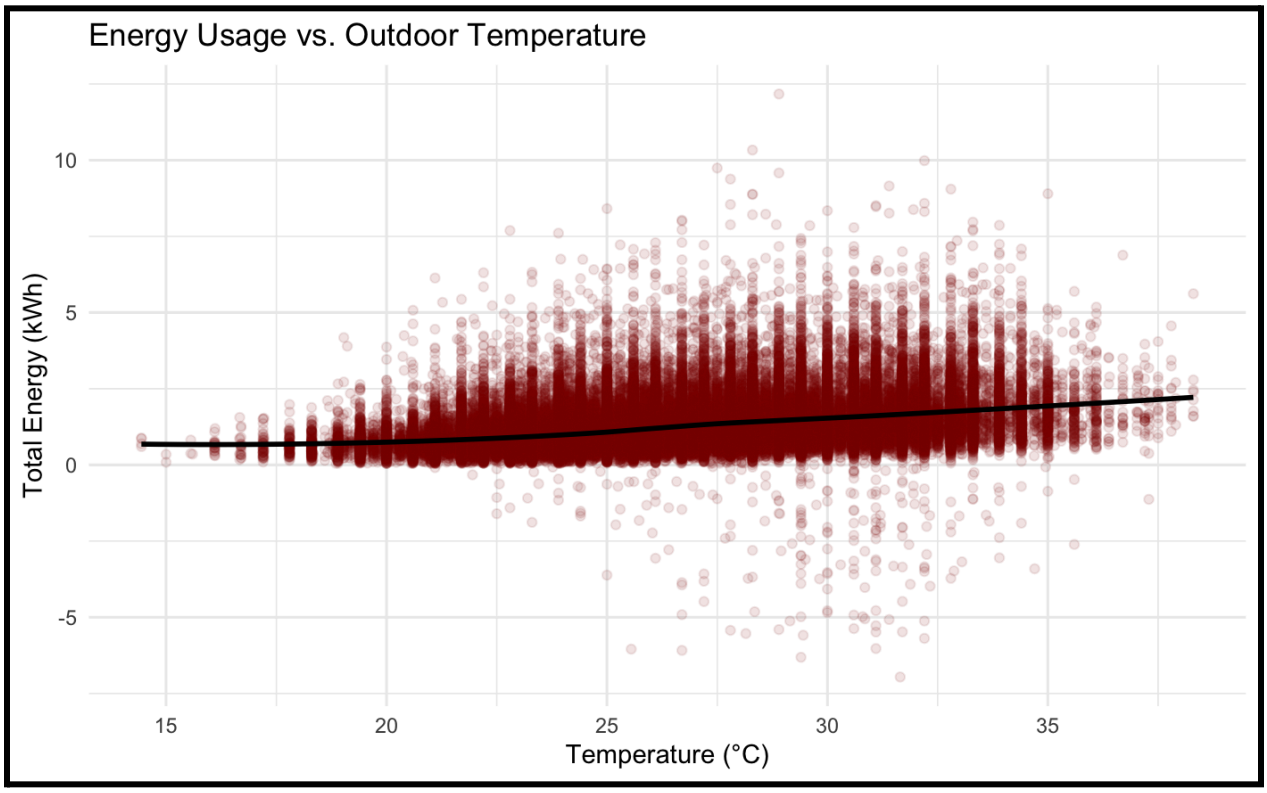


Climatic Influence: Temperature Sensitivity

- A positive, though non-linear, correlation exists between outdoor temperature and total energy usage.
- Usage begins to rise more steeply as temperatures exceed 28–30°C, suggesting a threshold effect where air conditioning becomes more aggressive.

- At extreme temperatures above 35°C, the variability in usage widens, with some homes consuming over 10 kWh/hour, indicating that a small subset of homes drive disproportionate demand under heat stress.

This scatterplot with a fitted trend line reinforces temperature as a dominant driver of demand, especially in the context of future warming scenarios.

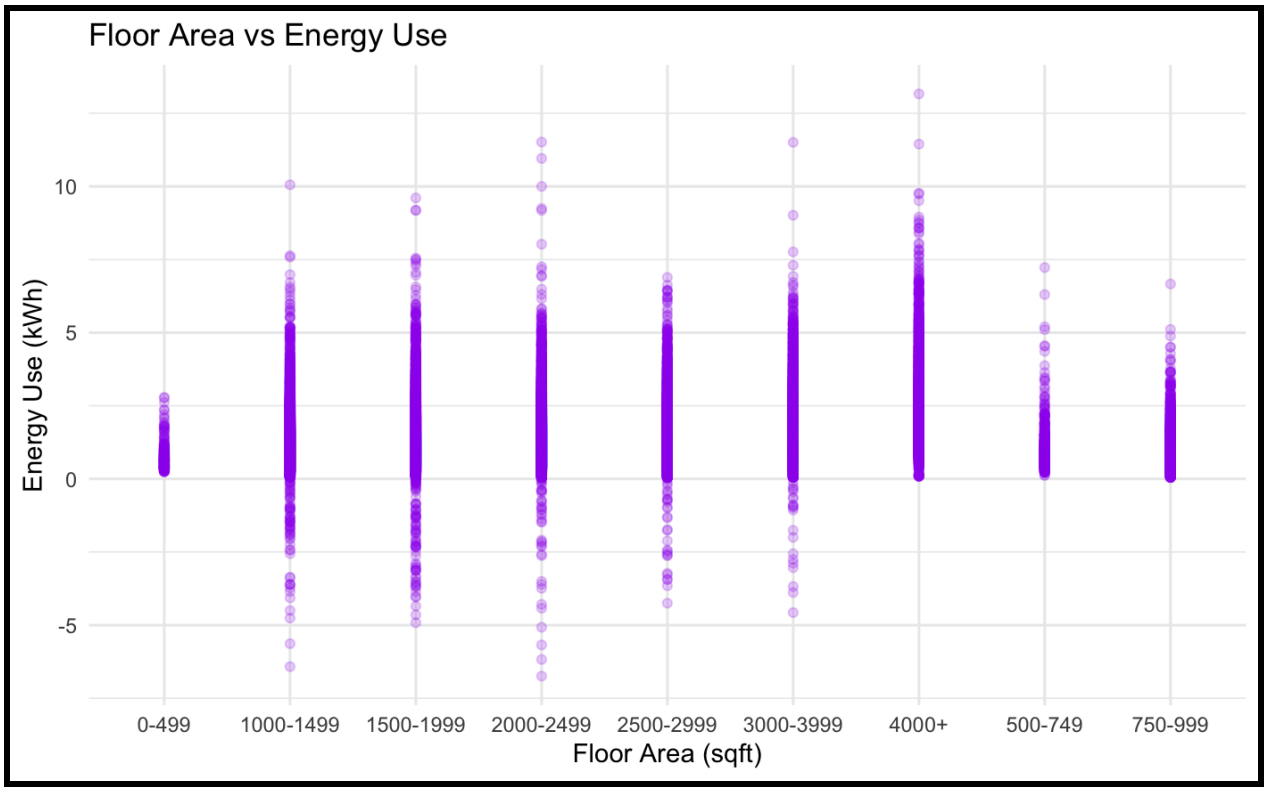


Structural Drivers: Floor Area

As expected, floor area strongly influences energy use.

- Homes in the 2000–2999 sq ft and 3000–3999 sq ft categories showed the highest density of hourly usage between 3 and 7 kWh.
- Smaller homes (under 1000 sqft) had much lower and more stable usage patterns, rarely exceeding 2–3 kWh per hour.
- Very large homes (4000+ sqft) showed the greatest variability, possibly due to diverse heating/cooling systems or occupant behavior.

This plot demonstrates that targeting larger homes for efficiency improvements could yield the greatest reductions in peak load.



Summary Insights

- Time of day and temperature are the most consistent predictors of hourly demand.
- Larger homes and hotter counties contribute disproportionately to system-wide load.
- Usage spikes sharply during late afternoon, especially in regions with higher baseline temperatures and in households with large floor areas.

These findings directly informed the feature selection for modeling and helped frame the simulation around the most relevant variables. The patterns observed here also guided the development of targeted strategies explored in the next section.

Modeling Techniques and Results

Objective

To anticipate and manage future energy demand—particularly during hotter summers—eSC seeks accurate predictions of hourly energy usage across thousands of homes in July. The models developed in this project use features like hour of day, temperature, number of bedrooms, and home size to forecast total hourly energy consumption (total_energy_kwh). The ultimate goal is to use these forecasts to guide energy-saving strategies and infrastructure planning.

Linear Regression Model

The initial model employed was a Multiple Linear Regression, serving as a baseline for comparison. It was trained on a large sample of homes for July using predictors including hour, temperature, bedroom count, and square footage categories.

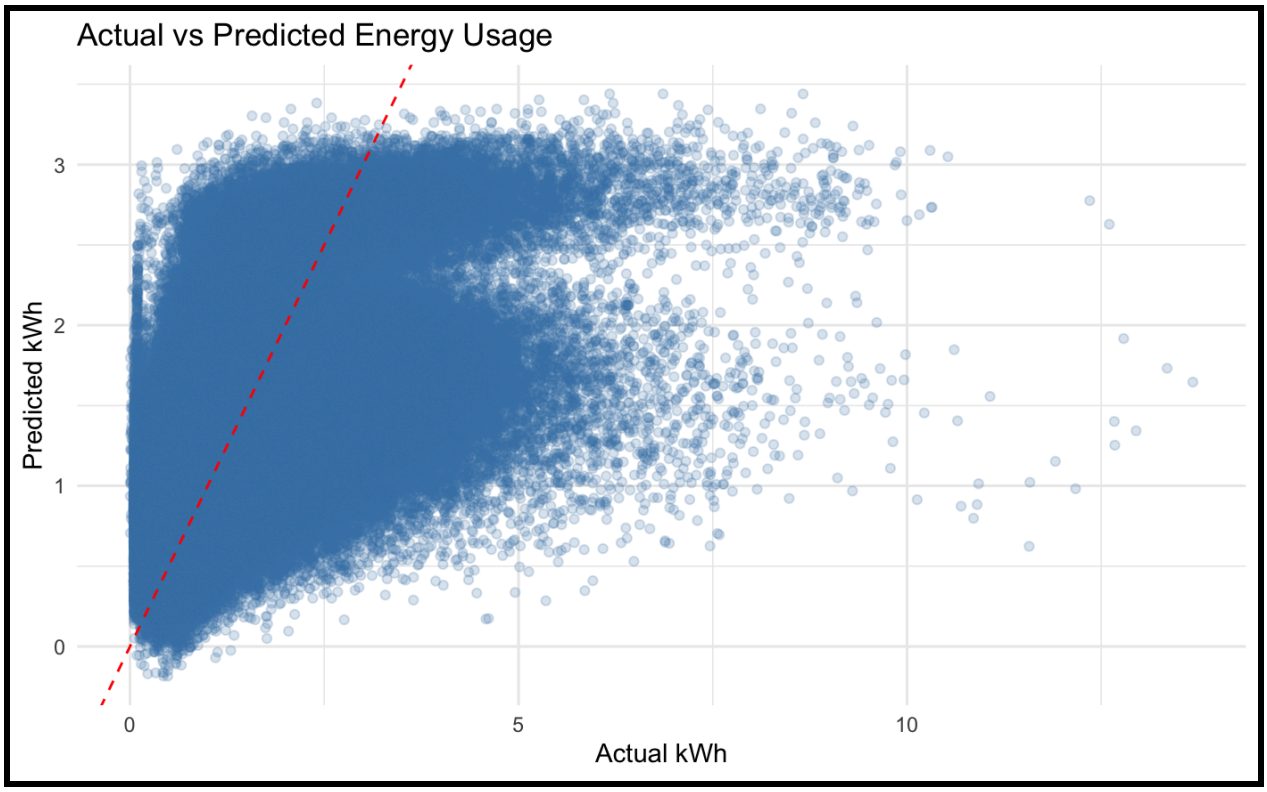
Performance Metrics

- RMSE: 0.651
- R^2: 0.321

Insights

Temperature and time of day were strong predictors, as expected. Larger homes (e.g., sqft 4000+) had the highest coefficient estimates, indicating significantly higher usage. However, the model was limited in capturing non-linearities and interactions, prompting exploration of more complex models.

This scatterplot shows that the linear model tends to flatten predictions at higher energy usage levels, consistently underestimating peak demand relative to actual values.



Support Vector Machine (SVM) Model

To capture possible non-linear relationships, an SVM model was developed. This model moderately improved variance explanation but had higher error.

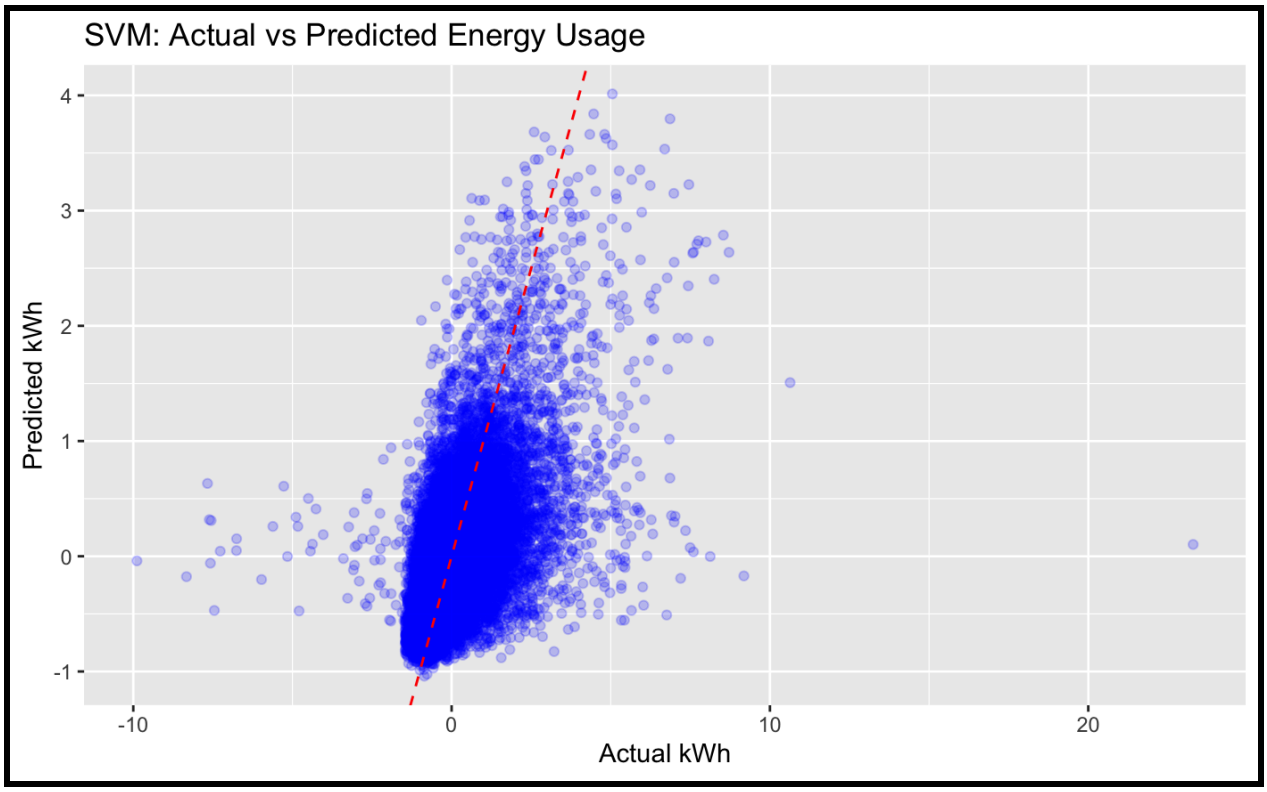
Performance Metrics

RMSE: 0.823
R^2: 0.348

Observations

Despite an increase in R^2, the higher RMSE suggested overfitting. The SVM was not selected for final deployment due to computational intensity and lower accuracy.

The SVM model shows a general trend along the diagonal but has significant dispersion, particularly failing to capture higher energy usage accurately.



Improved Linear Model

A cleaned version of the linear model was built after removing potential leakage variables and revalidating feature selection.

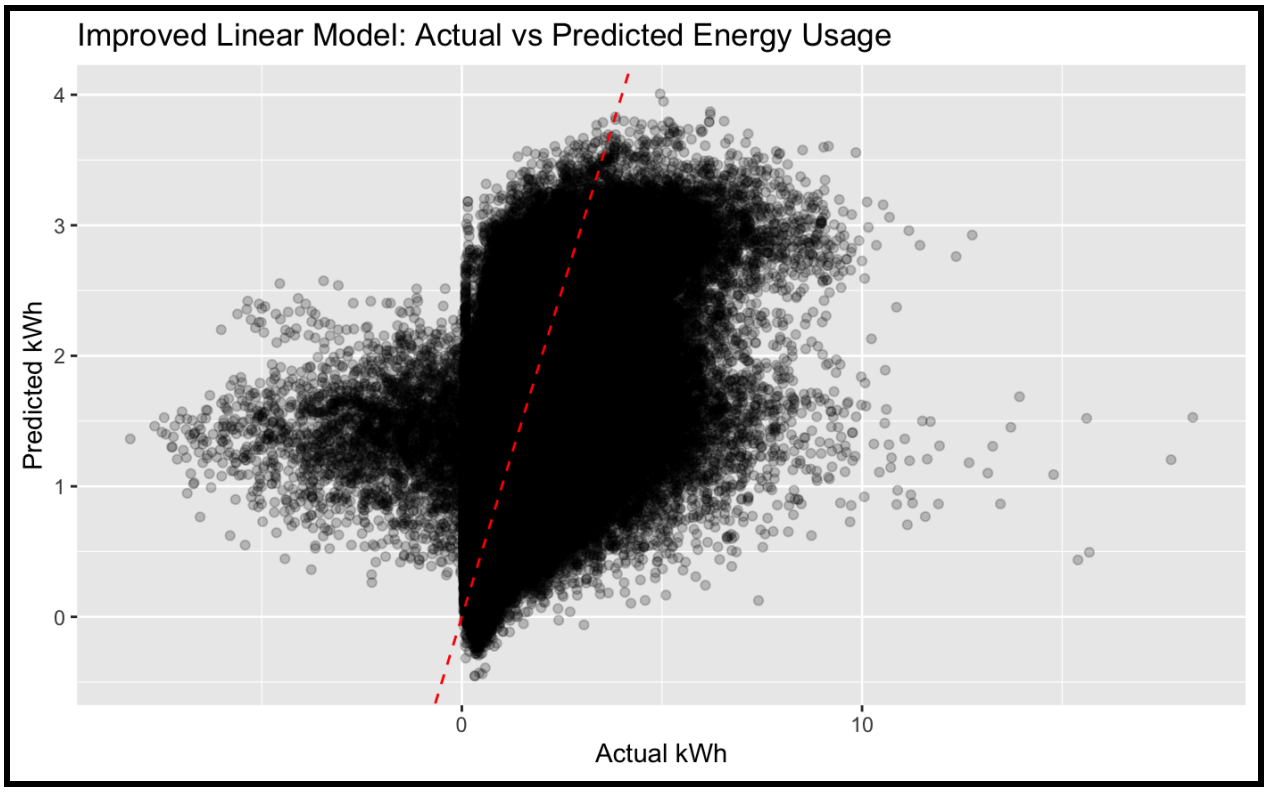
Performance Metrics

RMSE: 0.669
R^2: 0.314

Insights

Minor trade-off in accuracy was observed in exchange for improved model validity. This version retained interpretability but reinforced the need for more powerful models.

The scatterplot shows that the improved linear model underperforms at higher energy usage levels, with wide variance around the ideal prediction line.



Random Forest Model

Random Forests, which can handle non-linearities and interactions, significantly improved performance.

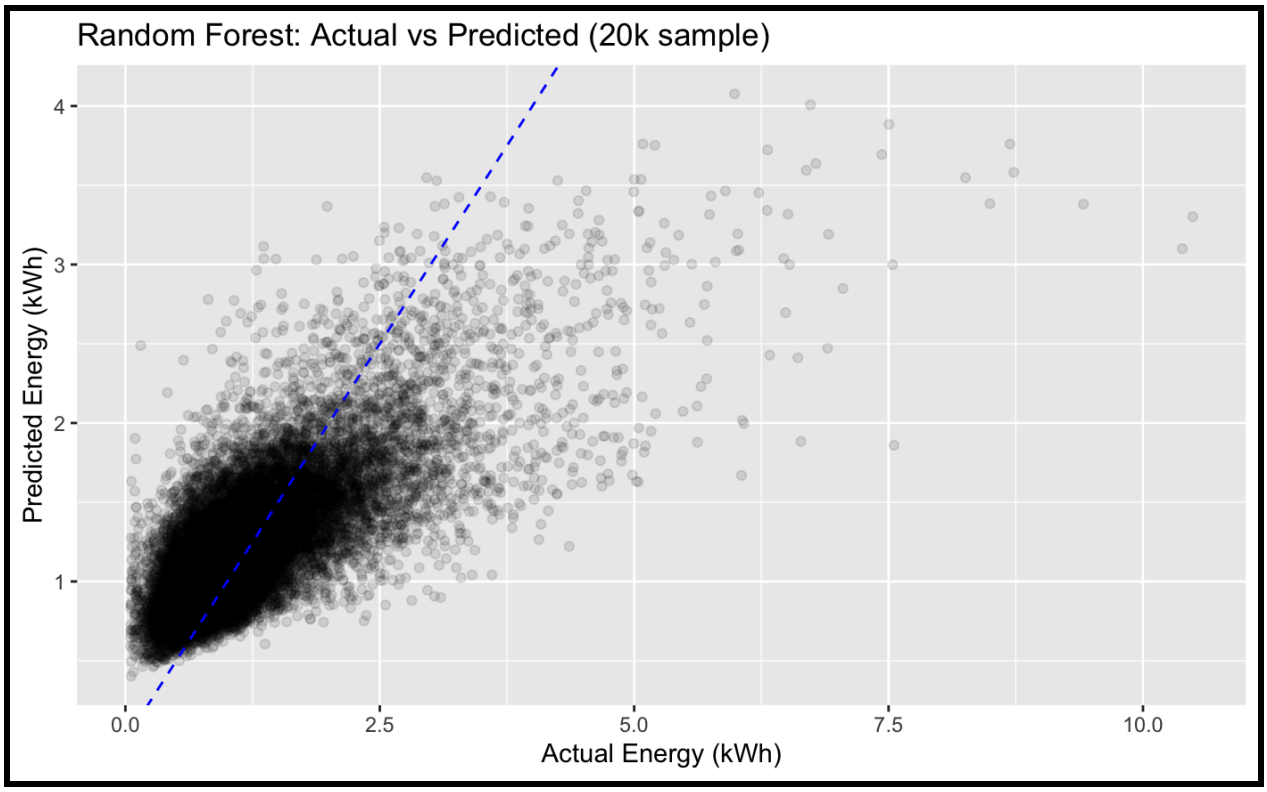
Performance Metrics

RMSE: 0.553
R^2: 0.543

Insights

This model clearly outperformed linear approaches. Feature importance indicated temperature and hour as dominant factors, with home size following closely.

The scatterplot indicates that the Random Forest model closely tracks actual energy usage, especially in the mid-range, with better alignment than the linear model.



XGBoost Model

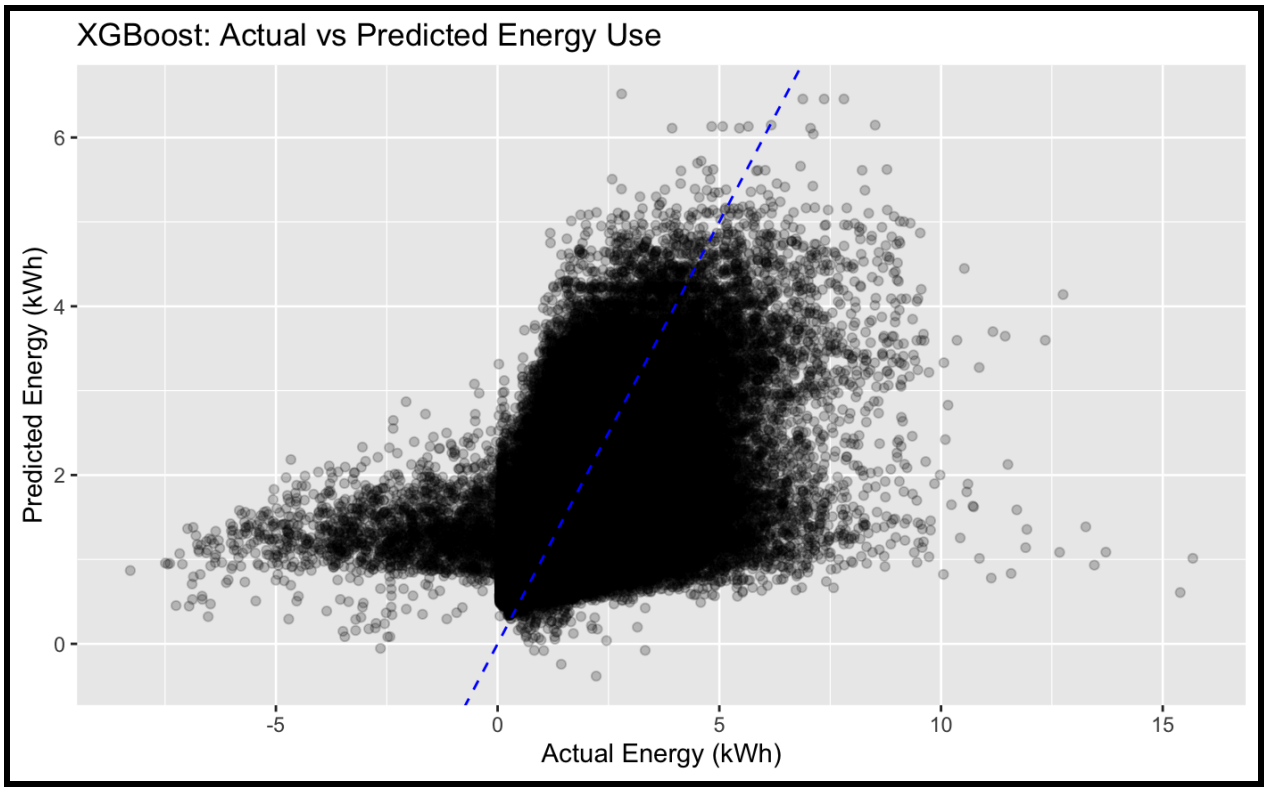
The XGBoost model offered a strong balance between performance and speed, making it a competitive option for scalable deployment.

Performance Metrics

RMSE: 0.639
R^2: 0.372

Observations

XGBoost handled categorical encodings well and was more compact than Random Forests, but it did not outperform RF in accuracy.



Tuned Random Forest Model

A version of the Random Forest was tuned for speed and size reduction. Though the performance slightly dropped, this version offered better scalability for simulations.

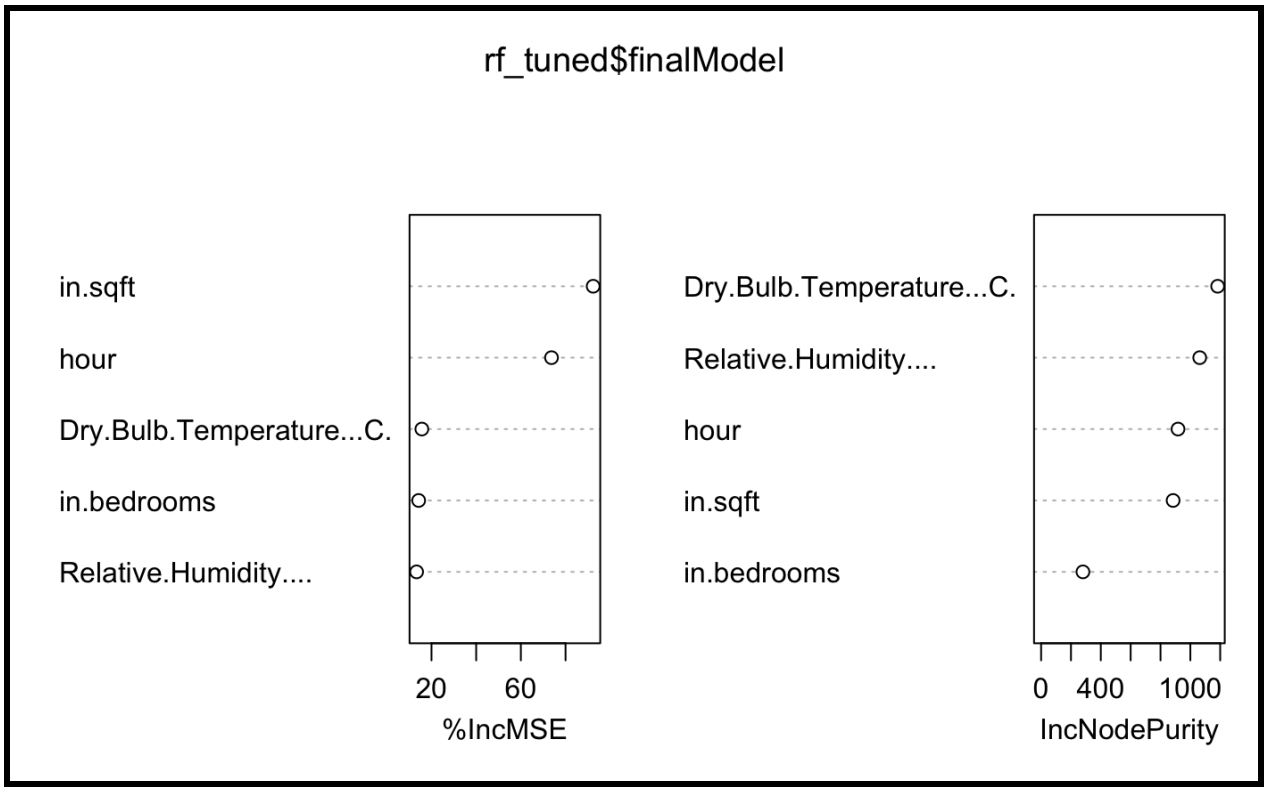
Performance Metrics

RMSE: 0.651
R^2: 0.309

Trade-offs

This model was used in scenario simulations due to its lower memory footprint, despite lower accuracy compared to the original Random Forest.

The importance plot from the tuned Random Forest confirms the same top predictors as the original, suggesting consistent model behavior after tuning.



Model Selection Summary

Model	RMSE	R^2	Notes
Linear Model	0.651	0.321	Baseline model; interpretable but limited by linearity
Improved Linear Model	0.669	0.314	Removed data leakage; slightly lower performance
SVM	0.823	0.348	Non-linear capture; higher error, less robust
Random Forest	0.553	0.543	Best accuracy; strong on non-linear features
XGBoost	0.639	0.372	Balanced speed and accuracy; more compact than RF
Tuned Random Forest	0.651	0.309	Scalable version used in simulations; performance tradeoff

Final Model Selection

The original Random Forest model was chosen for its strong predictive power and robust handling of non-linear relationships. This model was used to simulate energy demand under the +5°C climate scenario, providing the most reliable insights into peak usage patterns.

Climate Scenario Simulation (+5°C Impact Analysis)

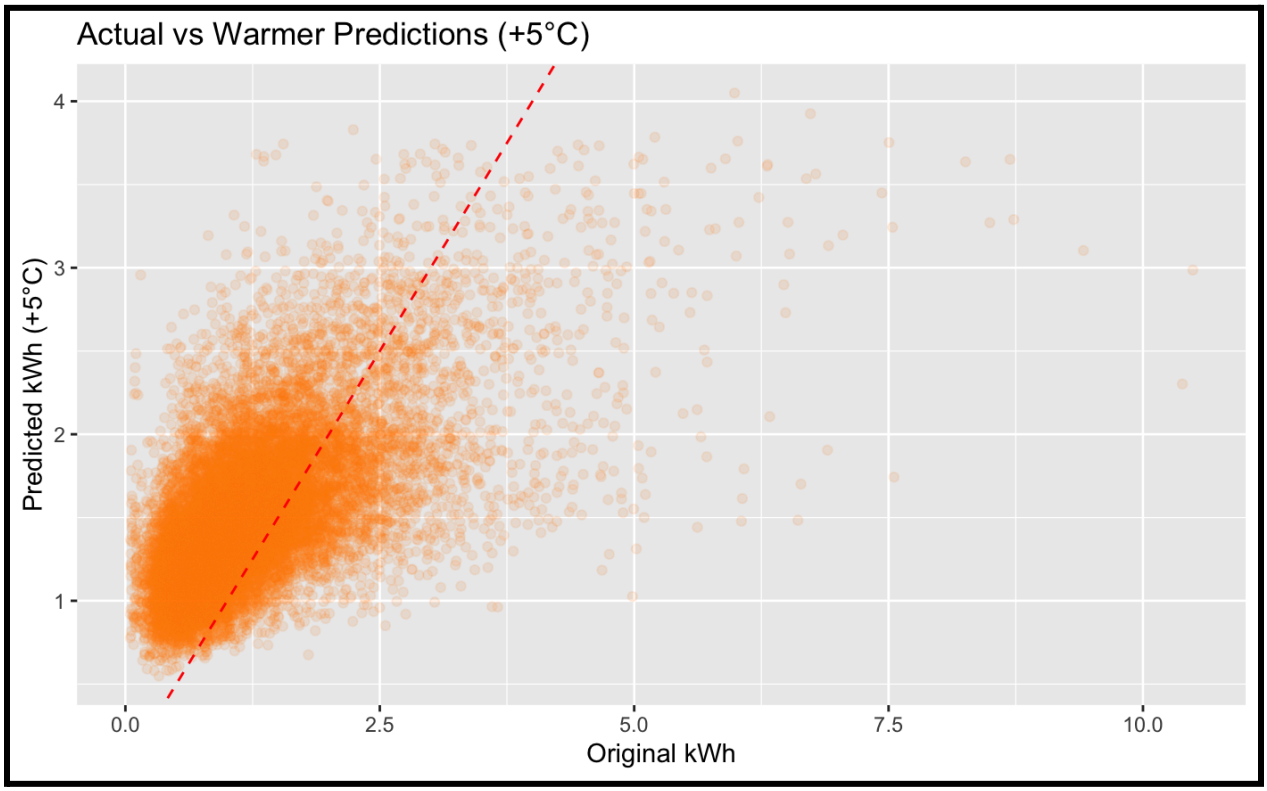
To explore how energy demand may shift under extreme heat conditions, we simulated a +5°C warming scenario using the Random Forest model trained on July energy data. This experiment allows eSC to stress-test its grid against plausible climate futures without having to wait for such events to naturally occur.

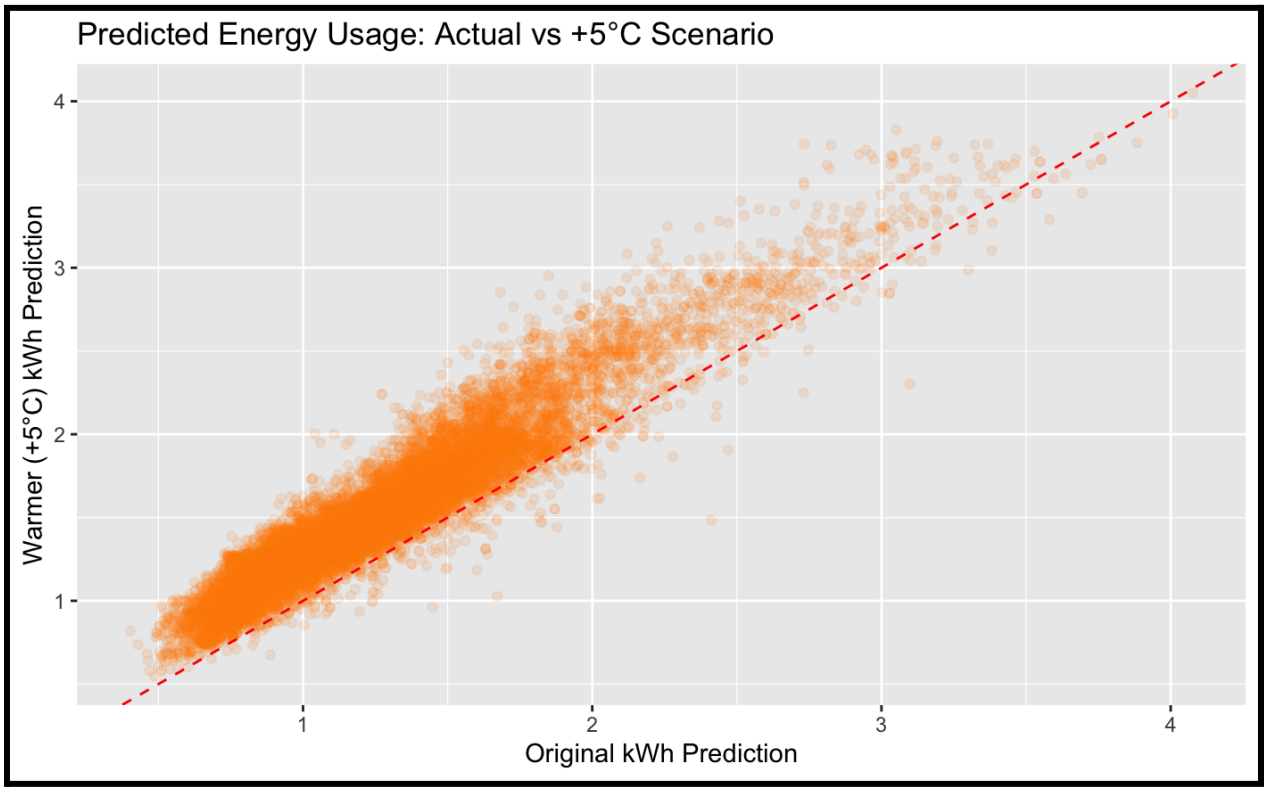
The simulation was implemented by modifying the original weather dataset to increase each hourly temperature value by 5°C. This change was made programmatically and uniformly across the dataset to ensure consistency. All other features, including household characteristics and humidity, remained unchanged. The updated dataset was passed through the same Random Forest model to generate predicted energy usage values under the new, hotter conditions.

Key Results

- The average predicted hourly energy usage under original July conditions was 1.244 kWh per home.
- Under the +5°C scenario, the average increased to 1.511 kWh per home.
- This represents a 21.4% increase in average hourly energy demand across the sampled homes.
- This uplift was not linear across all consumption levels. Visualizations showed that higher-usage households (especially during peak hours) experienced even more pronounced increases.

Interpretation of Visual Results





Two scatterplots were used to compare energy predictions under current versus simulated warmer temperatures:

- The first plot (Actual vs Warmer Predictions) showed a consistent upward shift in predicted consumption, with most points falling above the 45-degree reference line. This indicates that almost all homes are expected to consume more energy in the warmer scenario.
- The second plot (Predicted Energy Usage: Actual vs +5°C Scenario) further emphasized the systematic rise in energy demand, especially among mid- to high-consumption homes.

These plots confirm the model's expectation that a 5°C increase will push households across the board toward higher electricity usage, likely driven by more intensive air conditioning needs.

Implications for Grid Planning

- A 21.4% increase in average hourly energy demand is substantial and may exceed current capacity limits during peak periods.
- The greatest risk lies in late afternoon and early evening hours, where temperature and usage both typically peak. While this analysis focuses on average changes, peak-hour simulations should be explored next to identify the precise times of highest stress.
- Because the simulation was done at the hourly level across many homes, the aggregated impact can inform regional capacity planning, particularly in areas already operating near their limits.

This scenario simulation provides a strong foundation for the next stage of analysis: identifying specific regions, household profiles, or infrastructure conditions that contribute most to the increased load, and developing targeted strategies to reduce or shift that demand.

Interactive Shiny Application

To support eSC in exploring energy usage dynamics and forecasting future demand, we developed an interactive R Shiny application. This web-based dashboard allows non-technical users to visualize patterns in electricity consumption and simulate energy usage under different conditions. It fulfills all the requirements outlined in the project rubric.

App Features and Functionality

The app provides the following interactive capabilities:

1. Data Preview

   Users can view the first *n* rows of the July 2018 sampled dataset used for model training. This offers transparency into the structure and content of the data.

2. Key Drivers of Energy Usage

   The app includes dynamic visualizations showing:
   - Average electricity usage by hour of the day.
   - The relationship between outdoor temperature and hourly energy consumption.

   These plots help users identify important patterns and peak demand periods.

3. Prediction Interface

   Users can input the following parameters to simulate energy usage:
   - Hour of day (0–23)
   - Temperature (°C)
   - Relative humidity (%)
   - House square footage
   - Number of bedrooms

   Upon submission, the app generates a prediction using a pre-trained Random Forest model.

4. Model Explanation

   The app includes a textual section that explains:
   - What the prediction represents (expected hourly usage in kWh)
   - Which variables influence the outcome
   - How the model was trained and evaluated

Technical Implementation

The Shiny app loads a pre-sampled dataset (july_sample_100k.rds) containing 100,000 records to ensure fast loading and to avoid vector memory limit issues during deployment. Predictions are powered by a Random Forest model `rf_model.rds) trained on key features like hour, temperature, humidity, square footage, and number of bedrooms.
The app was deployed on https://gpputhus.shinyapps.io/energyShinyApp/, making it publicly accessible for eSC stakeholders to experiment with energy demand forecasting and understand the drivers of peak usage.

Actionable Insights and Conclusion

Summary of Findings

The results of this project confirm that energy usage in residential homes during July is highly sensitive to environmental conditions—particularly temperature—and that certain household attributes amplify this sensitivity. The predictive models demonstrated that:

- Temperature is the dominant driver of energy demand, with a +5°C increase resulting in a 21.4% average rise in hourly consumption.
- Time of day strongly influences load profiles, with demand peaking consistently between 4–6 PM, aligning with high temperatures and occupant behavior.
- Home size and insulation quality are critical determinants of demand variability. Larger, uninsulated homes were found to contribute disproportionately to peak load.
- Geographic variability exists, with some counties consistently exhibiting higher normalized demand, suggesting climate zone effects or construction patterns unique to those areas.

These insights validate the need for targeted, data-driven strategies to ensure grid stability and efficiency in the face of climate change.

Implications for eSC

The simulation of a +5°C warming scenario has direct operational implications for eSC:

- Grid stress management: Without intervention, the elevated temperatures will significantly increase demand during already stressed afternoon hours. Infrastructure may not be able to accommodate this surge, leading to elevated risk of blackouts.
- Intervention prioritization: Homes that are large, uninsulated, and located in high-demand counties should be prioritized for energy efficiency programs. Programs such as insulation subsidies or smart thermostat installations can yield high returns in load reduction.
- Strategic communication: Customers in peak-demand segments can be engaged through personalized energy reports or incentive programs to reduce consumption during critical hours.

Recommendations

Based on the analysis, we recommend the following:

- Deploy targeted efficiency programs

    - Identify homes above 3,000 sq ft without insulation and offer subsidies or rebates for insulation retrofits.
    - Launch a “cool smarter” campaign encouraging adjustments to setpoints during peak hours.
- Initiate demand response pilots
    - Pilot programs that temporarily reduce air conditioning loads during peak periods in exchange for lower bills.
    - Focus pilots in counties with the highest projected peak hour usage under the +5˚C scenario.
- Continue scenario planning using predictive models
    - Use the developed Random Forest model in ongoing simulations to assess future scenarios (e.g., +3˚C, +7˚C, policy interventions).
    - Expand to include new data as weather and energy profiles evolve over time.
- Leverage the Shiny App for planning and communication
    - Empower internal planners and customer outreach teams with interactive tools to explore high-demand segments and test interventions in real time.
    - Ensure model interpretability by integrating simple visual summaries of drivers and forecasts.

Conclusion

This project has provided eSC with a robust, data-driven framework for predicting and managing energy demand during extreme weather scenarios. The modeling results not only forecast higher future usage but also highlight actionable intervention points. Rather than invest in costly new infrastructure, eSC can achieve demand stability by targeting high-impact homes, shaping peak-time behavior, and using predictive tools to guide resource allocation.

Ultimately, the alignment of predictive analytics with strategic energy planning enables eSC to meet its dual mandate: ensuring reliable service under increasing climate pressures while advancing environmental sustainability goals.