IST 687 Group 2 Project Report

Jonah Soos, Bhushan Jain, Ashish Ghaytadak, Aarya Mirkale, Pratham Vasani

Introduction:

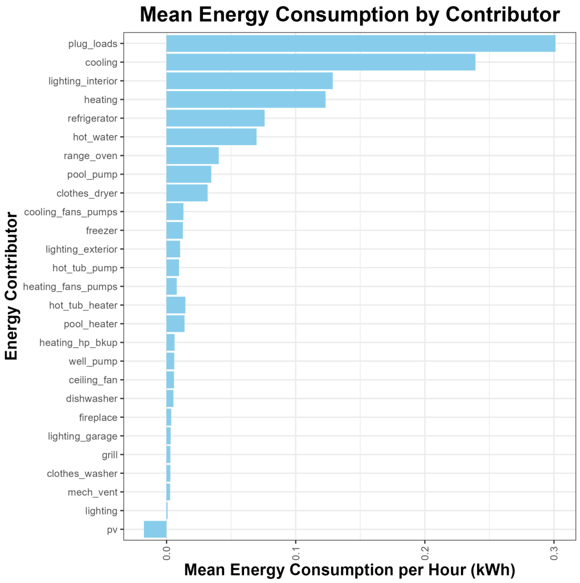
Being hired by the energy company eSC, our goal was to dive into the impact of global warming on the demand for energy within residential properties. eSC expressed worries that the following semester would overstress their current electrical grid, with many residential homes expected to increase power consumption due to rising temperatures causing blackouts. Building another power plant or other means to deliver more energy is not feasible, making data-driven initiatives to decrease energy consumption even more necessary. The focus month is July, as it is typically the month of highest energy usage.

Our goal for the project is to predict and manage energy demand during peak summer conditions by identifying actionable strategies for energy efficiency. We plan to integrate data, create predictive models, and provide proactive recommendations focused on appliance efficiency, user behavior, and weather impact on energy usage. This will result in data-driven insights and interactive tools for energy-saving strategies within peak demand conditions without increasing cost.

Methodology:

To address the goals at hand, we were provided with housing data for roughly 5,000 houses containing static information about each house to uniquely identify and describe it. We were also provided with information to acquire hourly energy consumption data for each house for the year 2018 and hourly weather data for each county of residence for the same year. The energy data consists of calibrated and validated energy usage, describing the usage of energy from many different sources for a specific house. The weather data consisted of temperature, humidity, wind, and precipitation information for each county within the housing data. These datasets were acquired by looping and appending through unique building and county keys and merged together by common columns of time and date.

Given this data, we ran into a major hurdle early on: the size. Considering it contained hourly information for an entire year, tens of millions of observations were recorded in total, making analyzing the entire dataset nearly impossible. To account for this, we first randomly sampled 500 households from the data to conduct a summary analysis, finding county, city, and time statistical trends. We understand that these may not be perfectly accurate to the entire distribution, but we felt that given a large random distribution, important trends would be recognized and representative of the entire data. For modeling purposes, we used all houses but only the months of June to September, as the summer months were the focus of the analysis. This helped reduce the size of the data in our predictive models while still allowing the analysis to incorporate as much information as possible.



The next step in our analysis was preparing for the modeling process. Once our datasets were completed, we first looked at the feature selection. To decide what to include in our models, we first analyzed the energy dataset to find key contributors. Variables like plug_loads, cooling, lighting, heating, and other appliances were found to be the largest contributors to energy consumption when evaluating the mean hourly energy consumption. We then looked at the static house data and evaluated each column we felt had an intuitive impact on energy consumption that was supported by our primary energy analysis. Information like region, climate, cities, appliances, cooling and heating information, insulation, and house descriptive information were all further analyzed at a granular level to select the final variables. Variables we expected to have an impact, like plug loads or climate, did not show any variation among houses and were therefore removed. Other variables like insulation showed too much specific variation to be simplified for modeling purposes and likely would suffer from a lack of sample size among houses to be included as a categorical variable. Finally, many categorical variables were either simplified or converted to dummy variables, including the presence of a dishwasher, appliance fuels, and appliance efficiencies. Once the modeling dataset was created, it totaled 35 variables and 12.6 million observations. The unit of observation was total energy consumption per hour, meaning all predictions are on the hourly level.

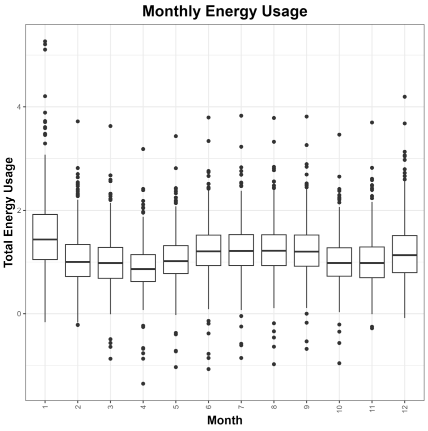
Modeling:

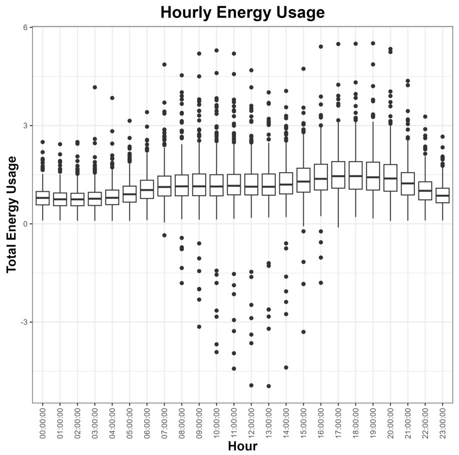
For the modeling process, three different model types were tested to find which could provide the most accurate actionable insights. All model types dropped any rows with NA values (totaling about 93,000 observations lost, less than 0.01% of the observations). The first model type was a linear OLS model, used as a baseline and to validate decisions prior. Before running the model, partial correlations were run between each variable to test for multicollinearity, finding that all appliance efficiency variables were perfectly correlated with one another and could be combined into one singular variable. Through running the initial regression, singularity was found within heating type and heating fuel, leading to both being removed. Income was also removed as no actionable insights could be made from the categorical type with too many brackets of too few sample sizes. The OLS model was found to explain 50.44% of the variation of the dependent variable, total energy consumption, with a root mean squared error (RMSE) of 0.5913.

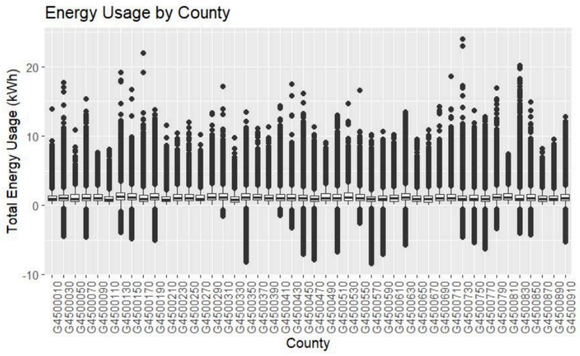
The next two models attempted to build on knowledge of the OLS model with an attempt to improve it. The first model run was a general additive model (GAM), which is similar to an OLS but removes the assumption of linearity and allows for parametric regression. With variables like temperature, it was expected to see some parametric relationships between the independent and dependent variables. It was found that 51.3% of the variance was explained by the model and predictions had an RMSE of 0.5866, a marginal improvement. The final model was a machine learning algorithm that uses decision trees to learn complex interactions between variables within the data, called XGBoost. This showed to be marginally the most accurate model; however, the black-box nature of the predictions makes creating actionable insights and the ability to point to specific coefficients representing expected changes impossible, making the model not the best option for this context. Given this conclusion, the GAM model was used for predictions as well as the summary analysis.

Results:

The first results we found were using the sample data to analyze specific trends.

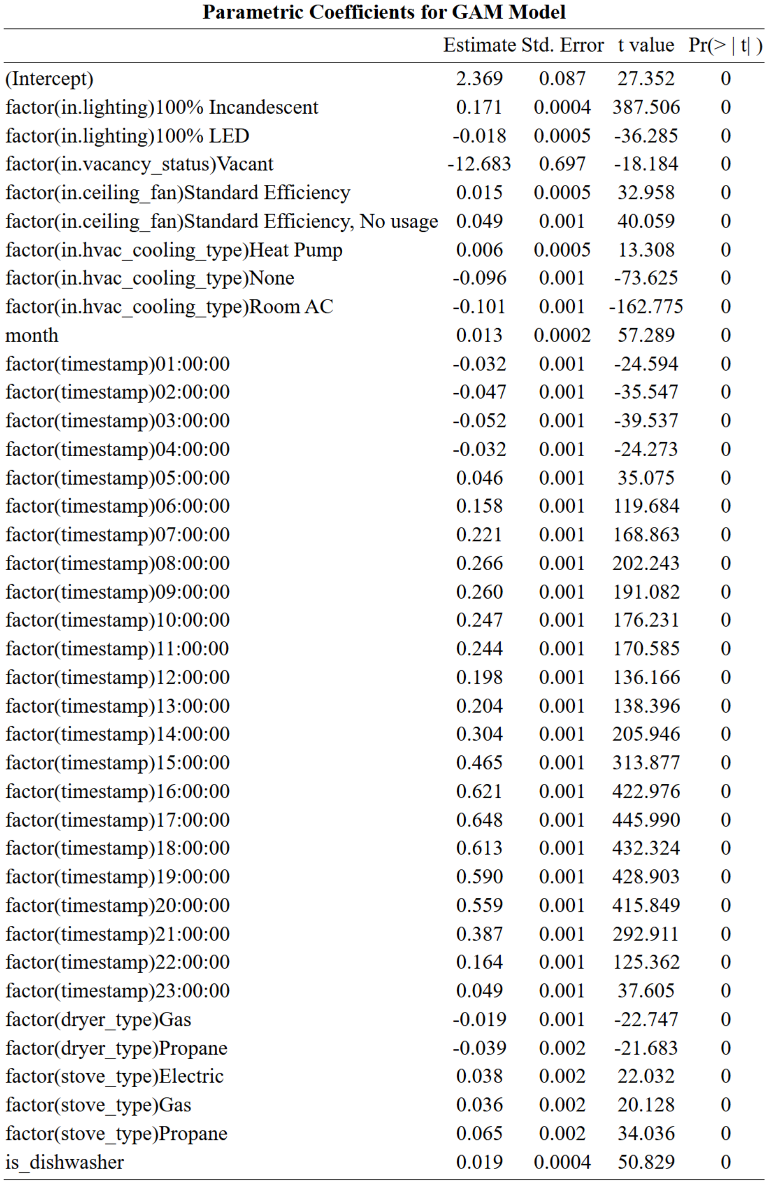






The first graph analyzes mean total house usage per hour per month, reaffirming our decision to use June to September for modeling. The next graph show that peak hourly usage is typically between 5:00 and 8:00pm, an important fact to advertise to consumers encouraging them to not use heavy appliances (like dishwashers or dryers) during those hours as it may lead to overloads and blackouts. The final one looks at usage by county, where the counties with the highest means represent urban areas that may be heavily affected by more houses in closer proximities with more residents.

Next, we can utilize our GAM model to find specific statistically backed coefficients that give actionable insights to customers. The first thing to notice is that the far-right column containing the p-value of each variable is 0 for each variable. This means that each coefficient is statistically significant, and evidence supports each conclusion we take away from this.



Parametric Coefficients for GAM Model

	Estimate	Std. Error	t value	Pr(> | t|)
(Intercept)	2.369	0.087	27.352	0
factor(in.lighting)100% Incandescent	0.171	0.0004	387.506	0
factor(in.lighting)100% LED	-0.018	0.0005	-36.285	0
factor(in.vacancy_status)Vacant	-12.683	0.697	-18.184	0
factor(in.ceiling_fan)Standard Efficiency	0.015	0.0005	32.958	0
factor(in.ceiling_fan)Standard Efficiency, No usage	0.049	0.001	40.059	0
factor(in.hvac_cooling_type)Heat Pump	0.006	0.0005	13.308	0
factor(in.hvac_cooling_type)None	-0.096	0.001	-73.625	0
factor(in.hvac_cooling_type)Room AC	-0.101	0.001	-162.775	0
month	0.013	0.0002	57.289	0
factor(timestamp)01:00:00	-0.032	0.001	-24.594	0
factor(timestamp)02:00:00	-0.047	0.001	-35.547	0
factor(timestamp)03:00:00	-0.052	0.001	-39.537	0
factor(timestamp)04:00:00	-0.032	0.001	-24.273	0
factor(timestamp)05:00:00	0.046	0.001	35.075	0
factor(timestamp)06:00:00	0.158	0.001	119.684	0
factor(timestamp)07:00:00	0.221	0.001	168.863	0
factor(timestamp)08:00:00	0.266	0.001	202.243	0
factor(timestamp)09:00:00	0.260	0.001	191.082	0
factor(timestamp)10:00:00	0.247	0.001	176.231	0
factor(timestamp)11:00:00	0.244	0.001	170.585	0
factor(timestamp)12:00:00	0.198	0.001	136.166	0
factor(timestamp)13:00:00	0.204	0.001	138.396	0
factor(timestamp)14:00:00	0.304	0.001	205.946	0
factor(timestamp)15:00:00	0.465	0.001	313.877	0
factor(timestamp)16:00:00	0.621	0.001	422.976	0
factor(timestamp)17:00:00	0.648	0.001	445.990	0
factor(timestamp)18:00:00	0.613	0.001	432.324	0
factor(timestamp)19:00:00	0.590	0.001	428.903	0
factor(timestamp)20:00:00	0.559	0.001	415.849	0
factor(timestamp)21:00:00	0.387	0.001	292.911	0
factor(timestamp)22:00:00	0.164	0.001	125.362	0
factor(timestamp)23:00:00	0.049	0.001	37.605	0
factor(dryer_type)Gas	-0.019	0.001	-22.747	0
factor(dryer_type)Propane	-0.039	0.002	-21.683	0
factor(stove_type)Electric	0.038	0.002	22.032	0
factor(stove_type)Gas	0.036	0.002	20.128	0
factor(stove_type)Propane	0.065	0.002	34.036	0
is_dishwasher	0.019	0.0004	50.829	0

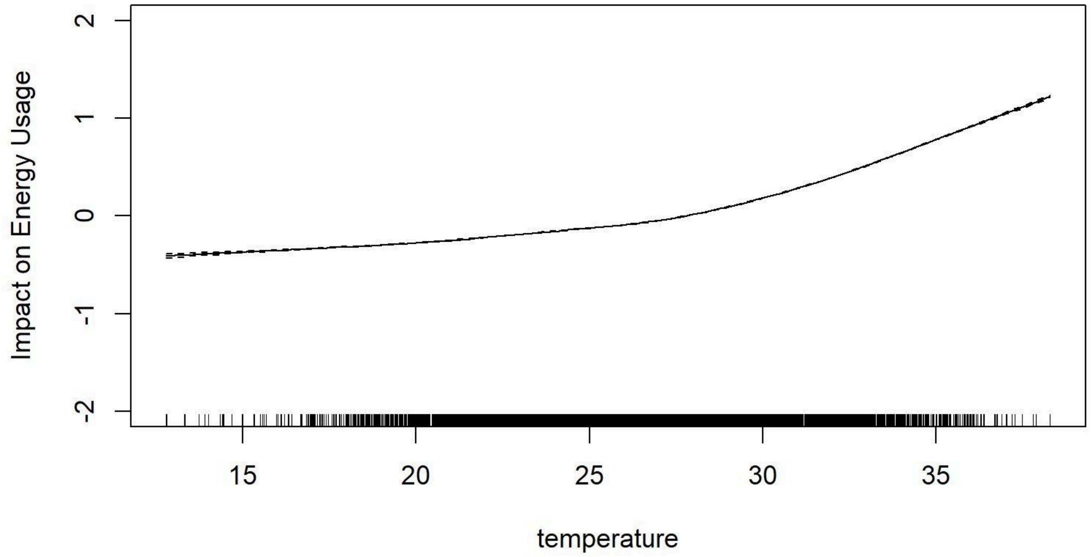
Our first intuitive conclusion, statistically backed, involves the timestamp variables. Each timestamp is being statistically compared to midnight, with the most negative being the lowest energy time periods (1:00 AM - 4:00 AM) and the most positive being the highest energy periods (5:00 PM - 8:00 PM). Using these coefficients, we also observe actionable insights, such as lighting types. Based on our model, it is expected that incandescent lighting utilizes 0.17 kWh more energy than CFL lighting, with LED lighting being another 0.017 kWh more efficient. This significant difference, with hourly means between 1–2 kWh, could significantly decrease energy demand during peak conditions.

This led us to suggest that eSC use public awareness campaigns to educate consumers on energy-saving behaviors, specifically promoting the switch from incandescent to LED lights. This would decrease energy consumption and energy bills while alleviating stress on the energy grid. Similar takeaways can be drawn regarding appliance types and fan usage. Houses with fans statistically use more energy (Standard Efficiency, No Usage having a coefficient of 0.049), which makes sense as they likely rely on more energy-intensive air conditioning units. Houses with room AC use about 0.1 kWh less energy per hour than those with central cooling, as they can turn on or off air conditioning in specific rooms based on use. Users with propane dryers and electric or gas stoves typically use the least amount of energy, whereas propane stove users and electric dryers are the most energy-intensive, albeit marginally so.

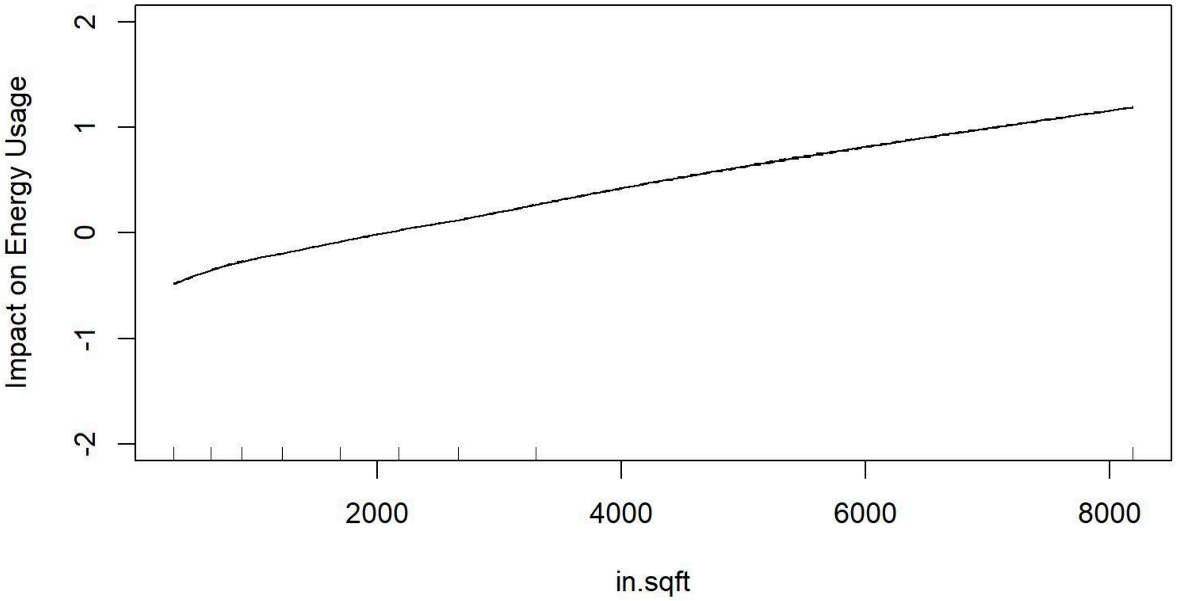
Smooth Terms for GAM Model

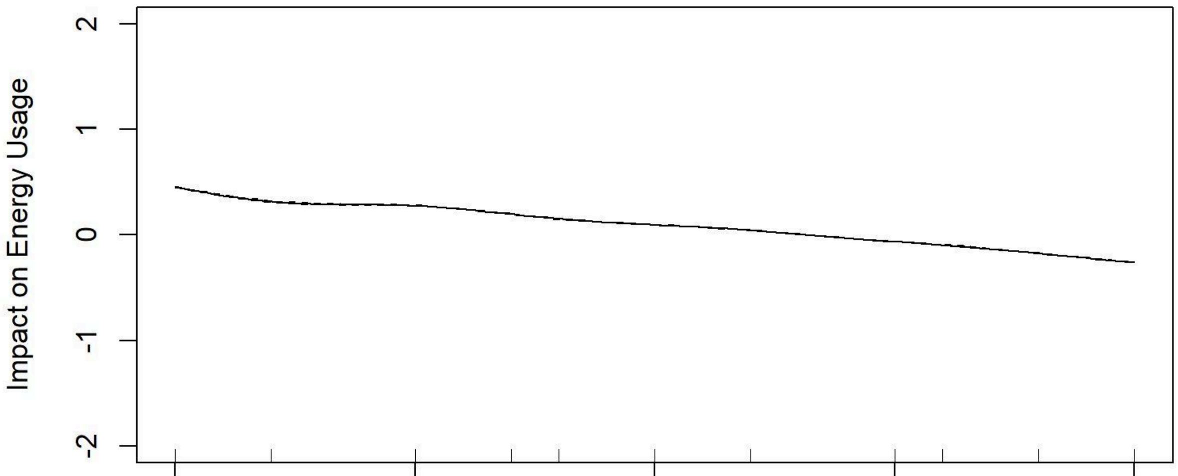
	edf	Ref.df	F	p-value
s(in.sqft)	6.997	7.000	210,582.700	0
s(in.bedrooms)	2.995	3.000	418.511	0
s(in.geometry_stories)	1.999	2.000	651.942	0
s(in.occupants)	7.990	8.000	41,679.150	0
s(in.cooling_setpoint)	8.997	9.000	54,311.080	0
s(in.heating_setpoint)	9.999	10.000	1,538.505	0
s(temperature)	8.560	8.937	48,302.710	0
s(humidity)	8.894	8.996	1,259.252	0
s(wind)	7.538	8.244	397.047	0
s(appliance_efficiency)	2.000	2	438,008.300	0

The next set of insights comes from the smoothed coefficients of the GAM model. These hold some of the greatest insights as graphically it can be seen the magnitude and change of the effects based on specific outputs. As seen in the table on the right, all variables are again statistically significant, allowing us to make real takeaways from the data and apply them within the contexts of the analysis. One of the most impactful parameters, as expected in temperature, exhibits a nearly linear relationship until reaching the peak summer heat above 27-28 degrees Celsius where energy consumption breaks its linear pattern and rapidly grows. With eSC predicting a 5-degree increase in temperature next summer, this validates their fears and shows the magnitude of the situation, where at about 35 degrees Celsius, the model expects nearly a full kWH of energy above the mean to be required.



The next impactful parameter was house size in square footage, showing along with temperature, the largest magnitude of energy impact based on the model. The relationship between square footage and energy consumption is again non-linear, exhibiting a trend of marginal returns, with at around 8000 square feet, eclipsing the 1 kWH threshold of predicted energy consumption impact per hour. This intuitively makes sense, larger houses require more energy to cool especially over the hot summer months, as the larger surface area requires more energy and air to maintain cooling setpoints. This intuitiveness reaffirms the model is accurately understanding relationships, and validates our analysis thus far. Another intuitive yet important impactor was the cooling setpoint. Despite not having as large of a magnitude of effect, it shows that setting the thermostat extremely low requires more energy to maintain it, especially over summer months. Some sort of smart thermostat regulations or subsidies for keeping it higher is more

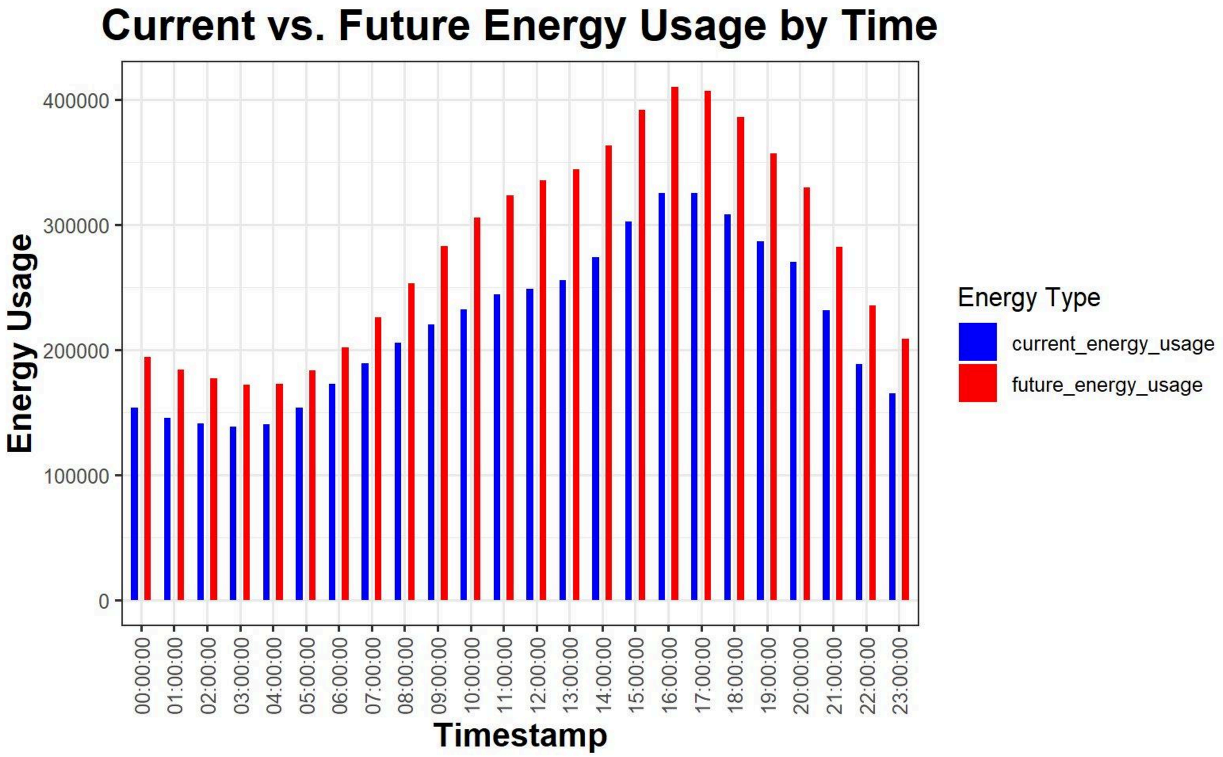




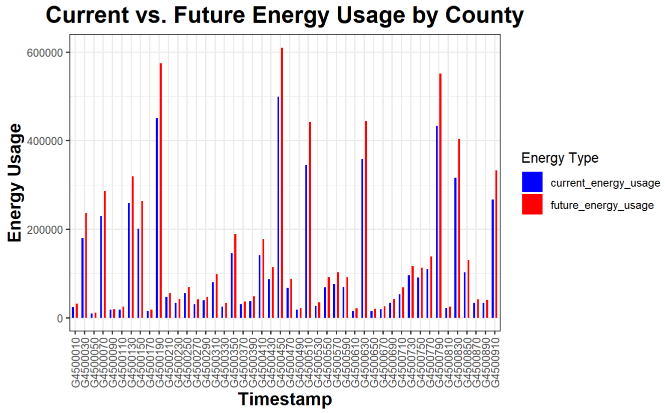
actionable ways eSC could encourage customers to use less energy potentially while adding in some fan usage talked about previously. Other relationships we found from our model were that more stories, bedrooms, and occupants lead to higher energy consumption, likely explaining similar effects as the square footage mentioned prior.

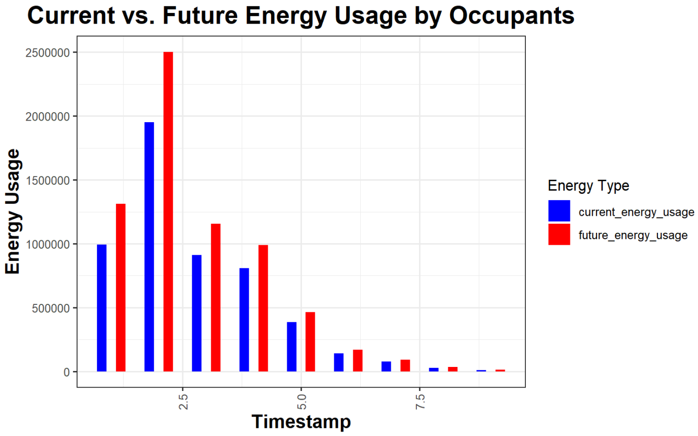
Predictive Analysis:

Our final task was to utilize the model and actionable insights discussed prior to explore what a future summer with 5-degree warmer temperatures would look like for energy consumption. It was found that the average peak demand increase in these conditions is predicted to be 26.48%, with a peak increase of 34.60% between the warmest times of the day, 11:00 am to 2:00 pm. The energy usage chart above shows the hour-by-hour total energy consumption expected during July, comparing the current to the forecasted, consistently showing those high red bars of large energy consumption increases if action is not taken. As expected, urban counties pointed out below are expected to see the largest change in energy demand while the percent change (difference in bar height) increases in larger family homes than smaller, with the total consumption seeing a large spike for 2-occupant homes, as they are most prominent in the larger metropolitan areas. Based on this final analysis, our third actionable insight is to attempt time-of-use pricing, encouraging off-peak energy usage for home appliances to avoid blackouts, especially in those heavily populated areas. If customers are charged more for using energy during peak hours, it may encourage them to utilize unnecessary appliances and goods during



off-peak hours. This would have to be explained with graphs and numbers so that clients could see when to use appliances and how that would impact cost, as would the thermostat subsidies and the public awareness campaigns for lighting.





Conclusions:

Overall we found statistical evidence that the eSC can use actionable insights to encourage customers to smartly use their appliances during offpeak hours and switch to more efficient energy to reduce their peak demand consumption. If eSC wants to interact further with the information, a shinyApp was created to show summary statistics of variables, a confusion matrix of predictions, and the predicted future energy consumption. In the future, more time to tune model parameters and utilize more months of data could likely improve these conclusions, as well as get more granular conclusions for specific house types, insulation, and cooling features, as well as further analyze yearly patterns and consistencies. Please contact us if any further analysis or explanations are necessary!