Data 102 Final Project

Canon Stringer, Omar Mejia, Joel Perez, and August Arneson

University of California, Berkeley

Data 102 - Fall 2024

Professor Strang and Professor Ramesh

Group 53

December 16, 2024

Contents

1 Data Overview

Our group decided to use dataset three, which focuses on electricity and carbon emissions forecasting. We employed real-time data from the US Energy Information Administration Hourly Electric Grid Monitor. This open API provides actual and forecasted electricity demand across various regions and states across the United States (excluding Alaska and Hawaii). The data encompasses an extensive amount of metrics including electricity demand, forecast, net generation by energy source, and carbon dioxide emissions.

The data is continuously collected every hour by the North American Electric Reliability Corporation, representing a census of power plants nationwide. We analyzed data from 2016-2023 for our first research question, and 2019 onwards in our second to ensure an accurate and contemporary analysis of electric demand trends as our focus is to examine more current up-to-date demand patterns. Additionally, instead of wrangling with a massive dataset, we only have to consider a subset.

Each row in the dataset represents hourly electricity demand for one of 13 regions in the lower 48 U.S. states. This fine-grained regional data allows us to observe trends at a regional level, but limits generalizability to specific states, counties, or cities. Consequently, our findings are regionalized and can not be more concise.

Because the dataset represents a census, there are minimal concerns regarding selection bias or convenience sampling. However, minor measurement errors can arise from factors like Daylight Savings Time, where one day each year includes an extra hour, while another day is shortened.

Some of the columns in the dataset contain missing data due to changes in reporting practices, such as the introduction or discontinuation of specific metrics. By focusing on a small range of data and filtering only specific columns, we eliminated many issues pertaining to missing data in the analysis.

Additionally, one of our research questions required utilizing historical weather data, which was obtained using an external dataset from the NOAA. The weather data contains important information that compliments the electricity demand dataset, such as variables describing temperature ranges, snowfall, precipitation, and wind speed which serve as useful confounding variables.

Each row represents a daily recording of the variables. This was able to be combined with regional specific California data grouped by the date. The result is a California region specific dataset where each row is a daily recording of the amount and types of energy sources consumed as well as the weather recorded. This additional dataset is a census, giving a partial concern for possible measurement error due to human mistake. Any missing data for precipitation, snowfall, and wind-speed were replaced with zero. We dropped average and minimum temperatures since we are focused on extreme weather, and dropped any rows missing the maximum temperature. This allows us to prioritize the maximum temperature as the outcome.

2 Research Questions

2.1 Question 1: Causal Inference

How do extreme heat events impact daily electricity demand in California?

2.2 Question 2: Prediction with GLMs and Nonparametric Models

Do time-based features predict electricity demand?

3 Prior Work

In our analysis, we investigated prior studies that help to give context to our research questions on electricity demand and its relationship with extreme heat events and time based features. Our key findings were:

Extreme Weather Events and Energy Systems

An analysis conducted by the National Renewable Energy Laboratory (NREL) investigated the relationship between extreme weather events and renewable energy. Notably, the study found that renewable energy sources, like wind and solar, are generally reliable for extreme weather crises, but are susceptible to natural phenomena limiting usage during events (cloudy skies, low wind). This is directly related to our first research question, which examines the causal effect of extreme heat events on electricity demand in California. The NREL focused on renewable energy, while our work quantifies the causal impact of extreme heat events. This is a good complement by providing granular insights into demand surges of California.

Temporal Patterns in Electricity Usage

The paper "Hourly electricity demand forecasting using Fourier analysis with feedback" by Yukseltan et al. developed a model to predict hourly electricity demand. The study utilized historical energy demand data and seasonal data to achieve a high accuracy. Unlike other models that include climate input, this model focused solely on using consumption and calendar information. While the model focuses on Turkey, it provides inspiration for how to incorporate temporal based features into our California predictive model. Our work furthers this analysis by comparing predictive modeling techniques.

4 Exploratory Data Analysis (EDA)

The primary goal of our EDA was to determine patterns and trends in electricity demand across the different regions in the US, while also considering factors such as hour, day of the week, and impacts of extreme heat events. We are looking for possible predictors of electricity demand and any unique anomalies we would need to address.

There seems to be a gradual incline in terms of electricity demand throughout the day (Figure 1). There is peak demand at 6pm and minimal demand at 3am, which makes sense and aligns with the traditional schedule of when Americans are awake. There does appear to be a trend in terms of how much demand there is at each hour of the day. This suggests that we could potentially predict electric demand. We will apply generalized linear models (GLMs) and non-parametric models to see if we find a good fit.

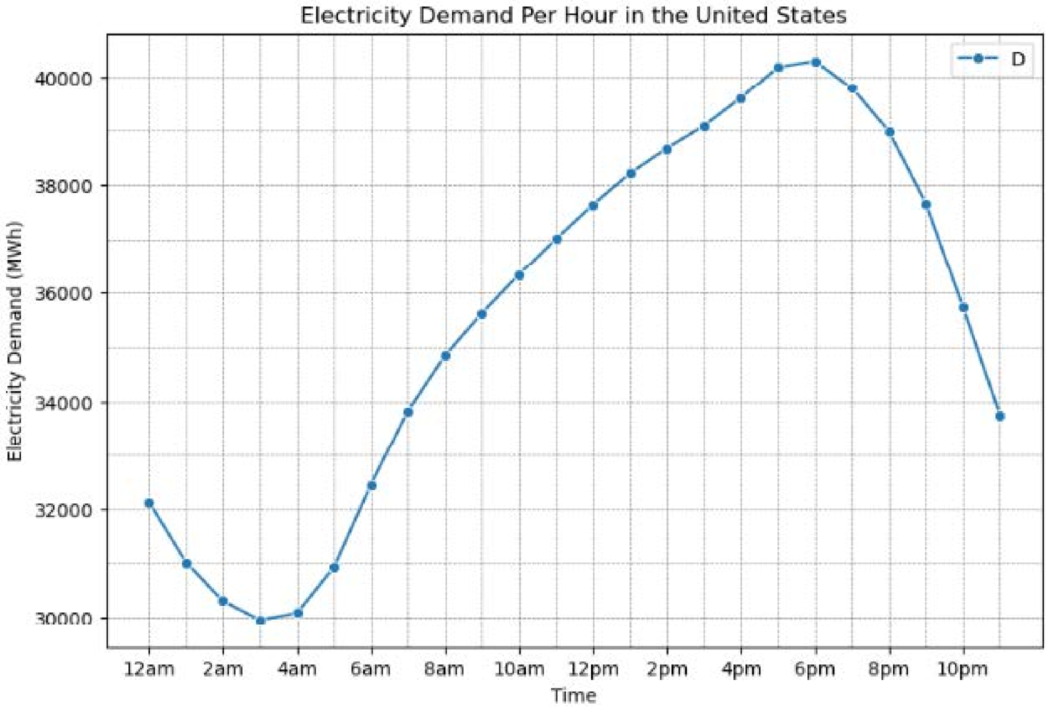


Figure 1: Electricity Demand Per Hour in the United States

Monthly trends are also relevant features (Figure 2). Since it gets hot in the summertime, people naturally start to persistently use air conditioning. Demand seems to be lowered during the Fall and Spring seasons (particularly October and April). Why? Maybe because most people do not necessarily need cooling or heating. It is an ideal temperature. Demand spikes a little in the winter, but not as much as we had anticipated. Timing in months seems to correlate to electric demand.

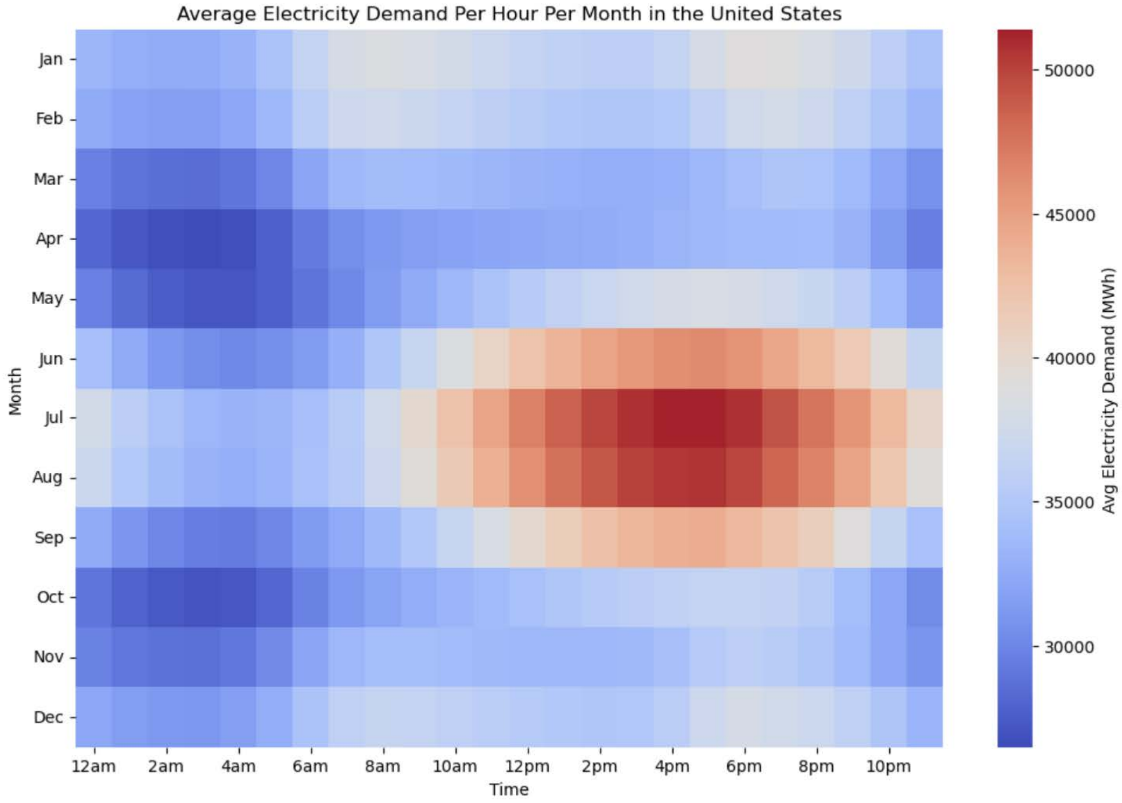


Figure 2: Average Electricity Demand Per Hour Per Month in the United States

There is also a notable difference between weekdays and weekends (Figure 3). There is much less demand for electricity on weekends than on weekdays. Why? Most people work their 8-5 jobs during that time. Many people return home around 5pm, and make dinner, consume media, or do a load of laundry, for example. Maybe people are getting up later on the weekends and not using as much electricity. We can take into account the days of the week when predicting the demand forecast because we believe there is a correlation with electricity demand.

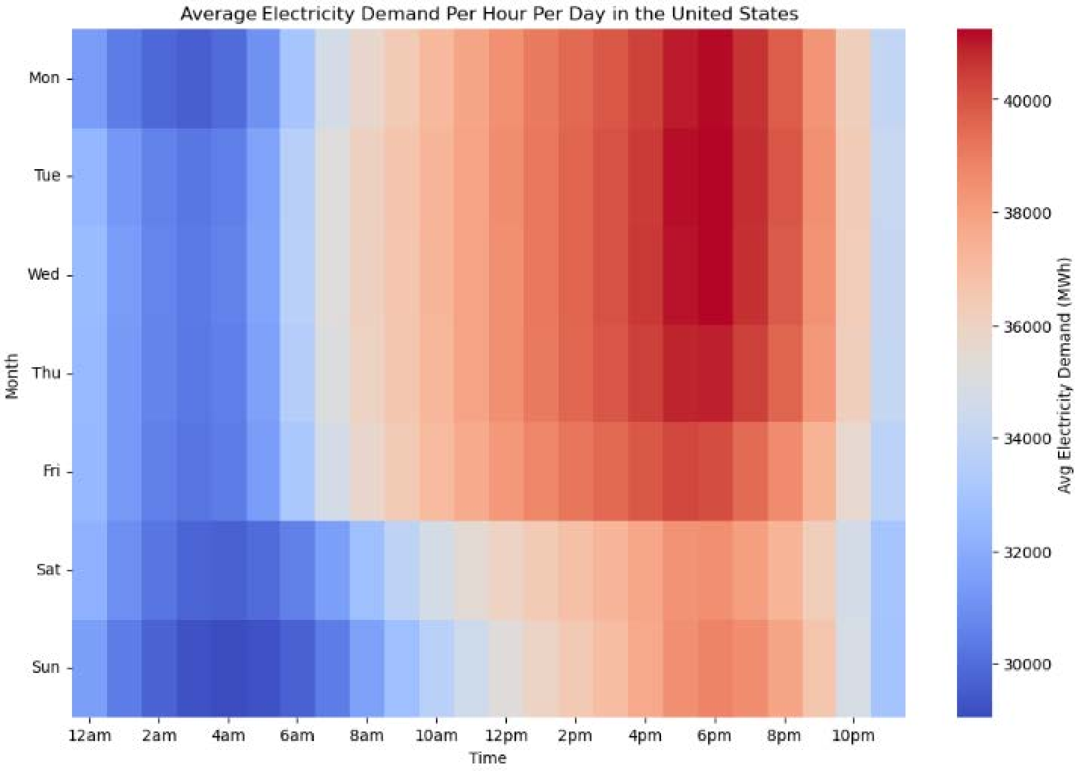


Figure 3: Average Electricity Demand Per Hour Per Day of the Week in the United States

We also questioned how extreme weather events impact daily electricity demand. To look into this, we analyzed the correlation between energy demand and different fuel sources (Figure 4). The goal was to identify any patterns in energy usage during extreme conditions. Our results showed that overall, coal, natural gas, and nuclear energy showed the highest correlation. Other energy sources showed weaker relationships. However, when comparing the top 10% of energy usage (extreme) to the bottom 90% (non-extreme), we were able to detect changes in correlation: coal decreased by 0.41, oil increased by 0.49, hydroelectric increased by 0.32, and wind decreased by 0.79. These results highlight the specific changes in energy sources during times of extreme demand, giving us an idea of the role of a specific fuel source that will be used during critical events.

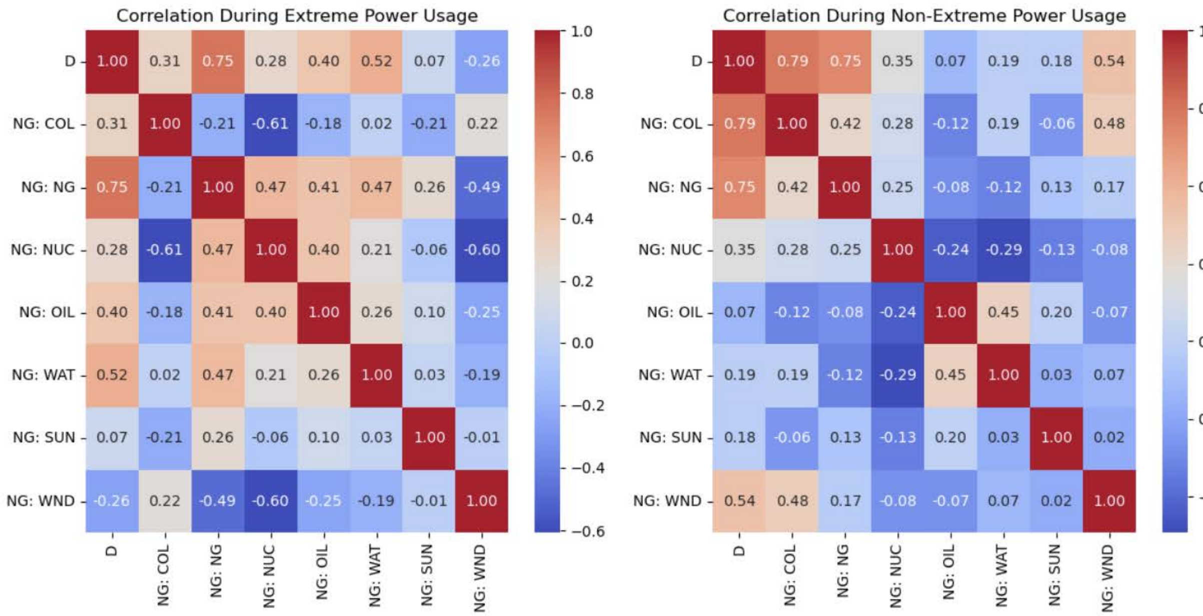


Figure 4: Correlation During Extreme and Non-Extreme Power Usage

We can see the highest correlation between maximum temperature and energy demand with 0.74. We also notice fairly strong negative associations with maximum temperature with precipitation (-0.43) and snowfall (-0.41). This helps highlight the present correlation heat waves can have with energy demand while also showing how strong environmental factors will affect heat waves.

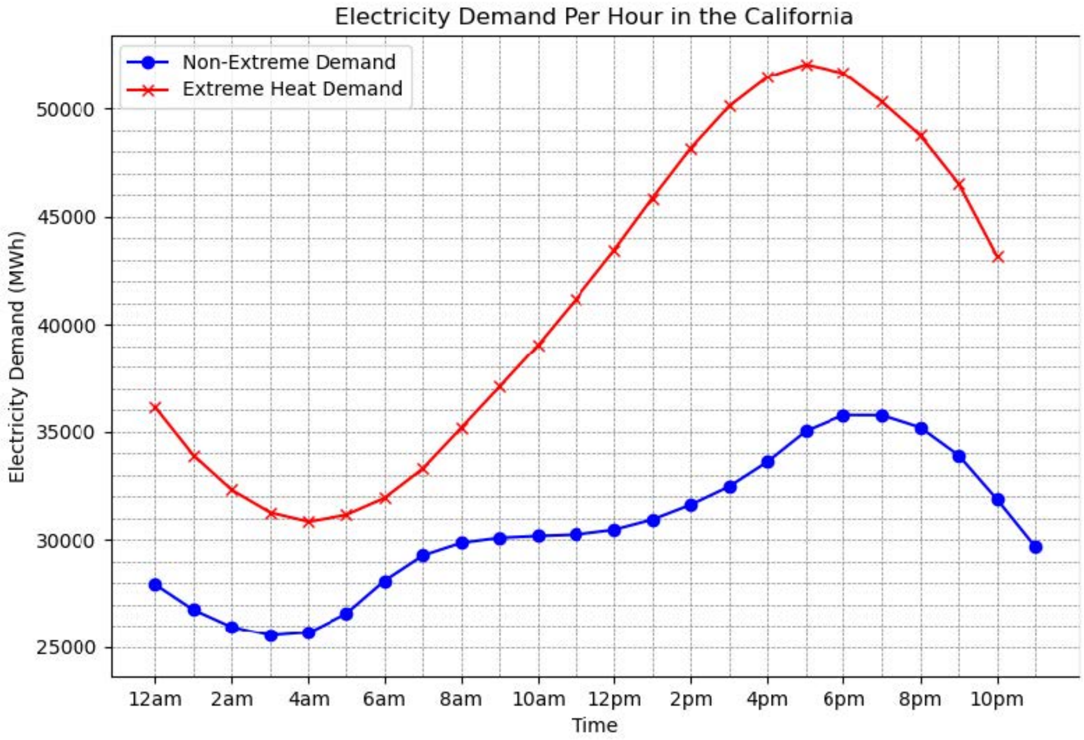


Figure 5: Electricity Demand Per Hour in California

This graph (Figure 5), in MWh, demonstrates that on average, at all times of the day electricity demand is significantly higher on days exhibiting ”extreme heat”. The biggest difference in the graph between the two is at around 4-6 pm, which also lines up with what are generally the peak electricity demand hours. We will determine to what extent this difference we see in this graph is causal.

5 Research Question 1: How do extreme heat events impact daily electricity demand in California? (Causal Inference)

5.1 Methods

The objective of this research question was to evaluate the causal effect of "extreme heat events" on electricity demand in California. The treatment was the occurrence of extreme heat in a particular day, with the control being days not experiencing extreme heat.

Matching techniques were utilized to estimate the causal effect of extreme heat events on electricity demand. Other techniques like inverse propensity weighting were considered and attempted but ruled out due to the rarity of these events occurring, as low-quality propensity scoring resulted in significant covariate imbalance.

The data granularity and units are such that each row represents one day's average temperature maximum across all weather stations and electricity demand in gigawatts per hour (GWh) in California from the years 2016-2023. These years were chosen as we believe them to be more relevant to modern discussions about energy usage, and there are few enough that there is less variance in the weather and energy data that could be accounted for by climate change and developments in energy related technology.

"Extreme heat events" were defined in line with the California Office of Environmental Health Hazard Assessment's definition of an extreme heat event - considering any daily maximum temperature (in degrees Fahrenheit) in the 95th percentile based on the respective month's historical average through the years 2016-2023 as a day of extreme heat (California Environmental Protection Agency).

Confounders examined through this matching process were the weekday, month, and whether that day exhibited significant rain, snow, or wind. "Significant" in the case of these weather events was determined by days exhibiting snowfall, rainfall or average wind in the 90th percentile of the data's historical average. C02 emission generation reported in the initial dataset was determined to be a colliding variable and was left out of the matching process due to it being affected by both the outcome and treatment. The confounder of "year" was left out of the matching process in order to prevent matches from being limited to the same year, enhancing data robustness. The relevant causal diagram is shown below:

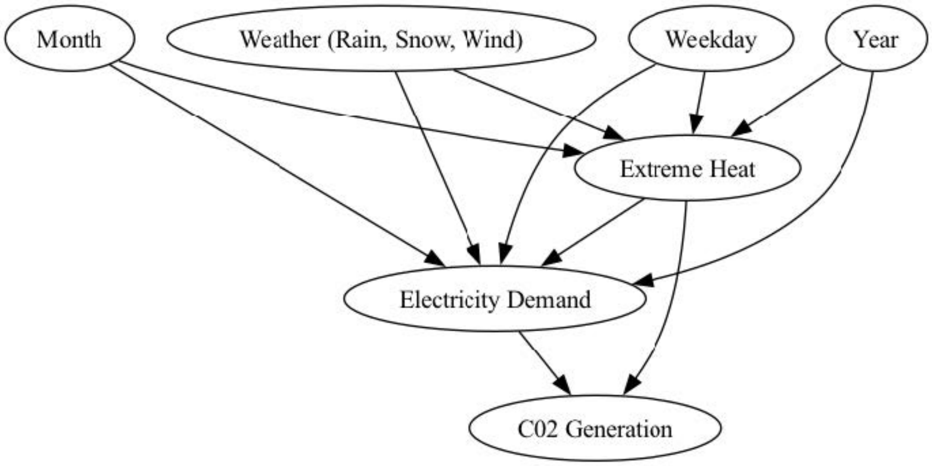


Figure 6: DAG: Casual Diagram

5.2 Results

After matching, 3,796 matches were found across these variables. Taking the average treatment effect on the treated (ATT) score on these groups yielded a value of approximately 95.8 GWh, making the causal effect statistically significant. We found the estimated causal effect of extreme heat on electricity demand in California is thus an increase of 95.8 GWh in demand. It is worth keeping in mind that confounding variables like year and potential social/behavioral variables were unaccounted for when interpreting the score. While excluding years from the matching process increased data robustness in one manner, years like 2020 that exhibited a change in electricity demand due to the COVID pandemic likely influenced this figure in an unknown manner.

Also important to note in the study is that the number of matches, and the outcome of the ATT score changes as we change our threshold for what is considered an “extreme heat” event. Below (Figure 7) is a bar chart showing how these numbers differ as we change our parameters:

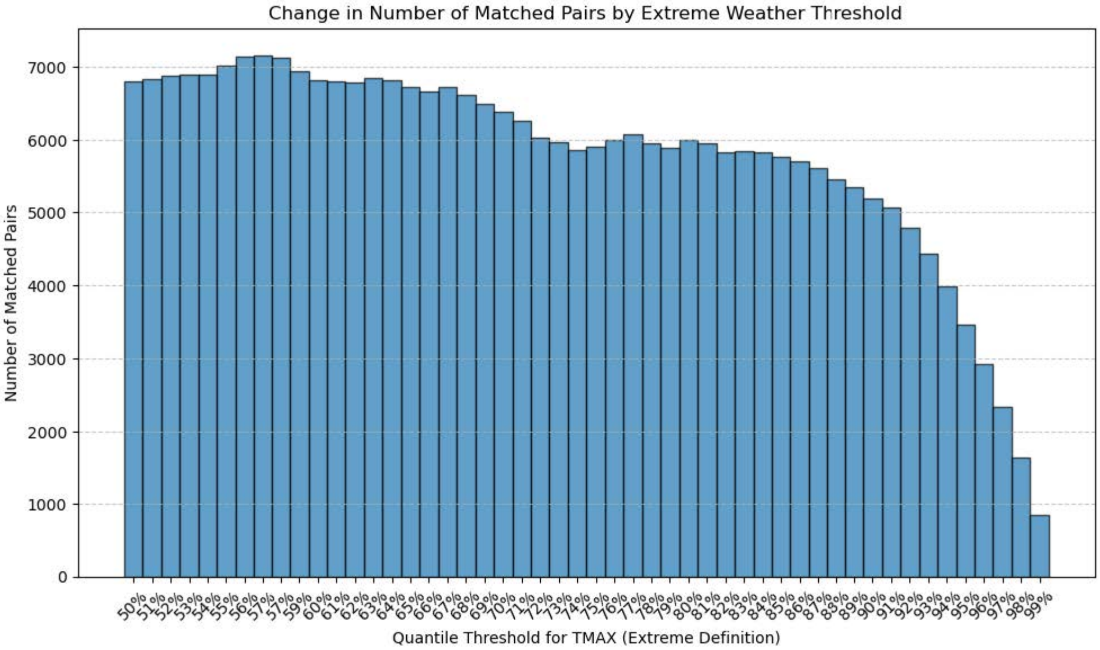


Figure 7: Change in Number of Matched Pairs by Extreme Weather Threshold

Evident from the chart above, there is a negative correlation between how many matches we get to calculate the ATT score and how high the quantile threshold is for extreme weather.

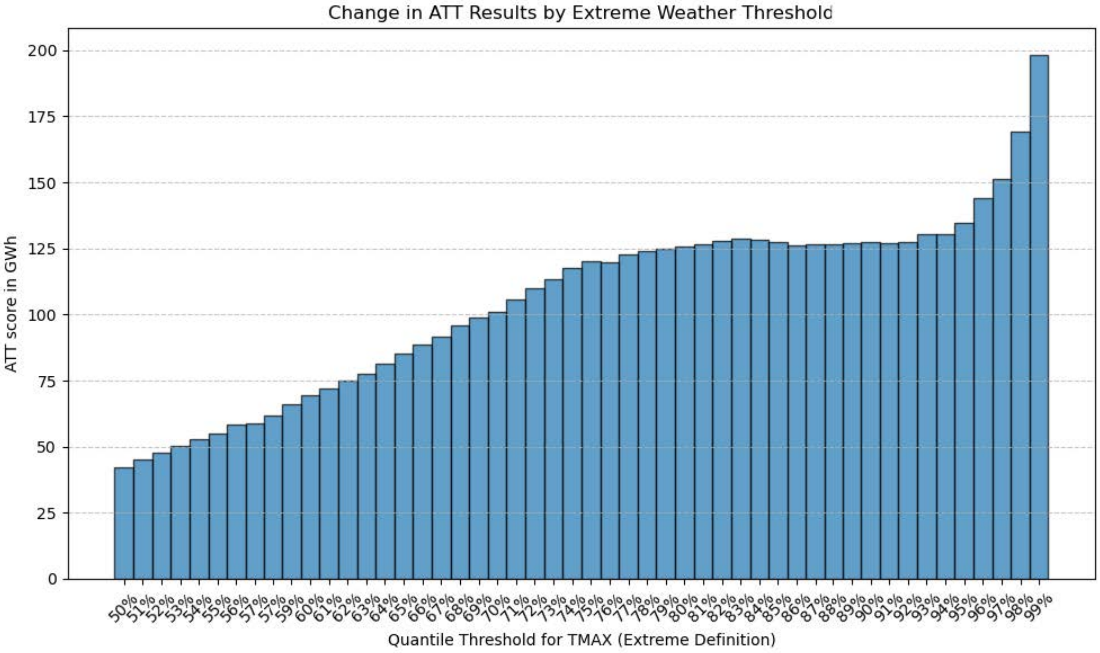


Figure 8: Change in ATT Results by Extreme Weather Threshold

In Figure 8, more importantly gives us insight into how varying levels of what we consider extreme heat impact the ATT output. The ATT value appears to steadily increase as the quantile threshold increases, before flattening from around 80% (equivalent to a maximum 85° Fahrenheit threshold) to 94% (91.1° F), and then exponentially increases from 95% (91.6° F) to 99% (95.7° F). The calculated difference in ATT scores between the 93rd quantile and 80th quantile came out to approximately 7.531 GWh, and by contrast this difference in score between the 99th and 93rd quantiles came out to approximately 68.129 GWh despite both pairs of quantiles being separated by about 5° F. Observing this is relevant as it shows that infrastructure to account for the demand of these 95%+ events should potentially be a focus, and also shows that there is little change in demand difference between the 80th and 93rd quantiles.

5.3 Discussion

While this outcome makes us confident that there is a positive causal relationship between electricity demand and extreme heat, largely due to an increase in cooling demand, limitations around this study leave us skeptical about the exact causal magnitude. As mentioned before, not all confounding variables were able to be accounted for in the matching process. Data regarding things like changes in energy efficiency over the period and major events impacting demand would enhance this estimate. Importantly too, the granularity of the data had to be kept state-wide due to the data available, giving us a high-level view of the energy needs in California overall but not on a county or city-wide level. Different regions in California utilize different energy infrastructure, experience different weather conditions, and may have varying social relationships to energy usage, so data tailored specific to a region would be more relevant in determining what changes that region should make to address the increase in demand.

What this data does tell us is that there is a concerning significant increase in electricity demand caused by extreme heat events in California. As climate change causes the frequency of these extreme heat events to increase, infrastructure to address this level of demand more frequently is going to be crucial. By pairing this along with information about how time-based features predict electricity-demand, we can optimize energy infrastructure to our needs.

6 Research Question 2: Do time-based features predict electricity demand? (Prediction with GLMs and Nonparametric Models)

6.1 Methods

Now let us transition to our next research question. Do time-based features predict electricity demand?

To enhance interpretability during EDA, we standardized the units in the "Demand" column, converting megawatts to gigawatts, where one gigawatt equals 1,000 megawatts. Given the magnitude of electricity demand, this scaling provides a clearer representation of the data.

In our analysis, we exclusively focused on time-based features, which consisted of hour, month, and day of the week.

Hourly electricity demand exhibits distinct patterns throughout the day. There is a visible, roughly linear increase in demand from the early morning hours until the evening, followed by a similarly linear decrease during the night. These fluctuations reflect your typical human activity patterns, such as increased energy usage during waking and working hours.

Additionally, we also look at the month of the year. Demand appears to peak during the summer months (June, July, and August) in the United States, in general. This trend likely corresponds to the increased use of air conditioning (AC) systems. AC units are known to be significant energy consumers. While this observation aligns with expected seasonal behavior, it's crucial to point out that we are making logical assumptions rather than direct evidence coming from the temperature dataset.

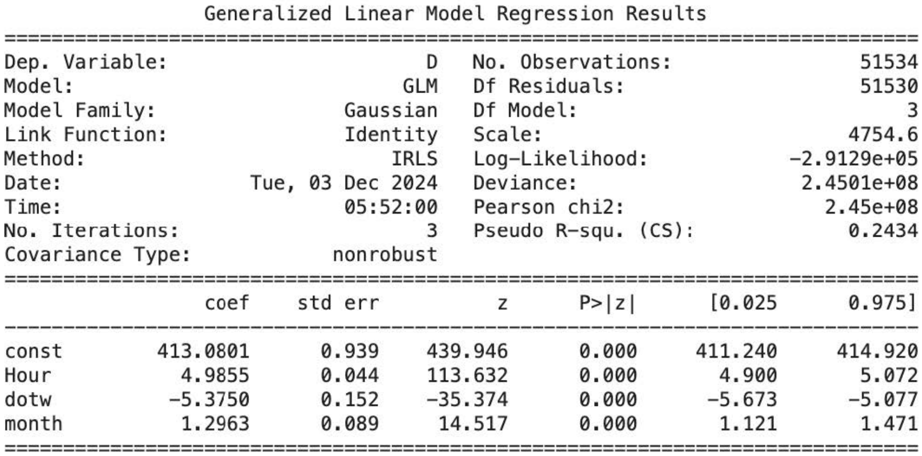
Analysis of the day of the week column revealed higher electricity demand during traditional work weekdays (Monday through Friday) and reduced demand on weekends (Saturday and Sunday). This finding aligns with the general expectation that commercial and industrial energy usage is higher on weekdays.

By isolating these particular time-based features, we are aiming to predict electricity demand. The following analysis provides a foundational understanding of how temporal patterns influence electricity demand.

6.2 Frequentist Model

We began our analysis by adopting a Frequentist perspective, treating the unknown coefficients of our model as fixed value. Unlike Bayesian approaches, we do not need to make any assumptions about prior distributions. Given the continuous nature of our outcome variable we are deploying a Gaussian linear regression model. Discrete and fixed models, such as Poisson or Negative Binomial regression, are not suitable for our context. Our model assumes that the response variable, electricity demand, is normally distributed around the linear predictors. Hence, we know that the likelihood function is Gaussian, the inverse link is identity, and the link function is also inverse. To evaluate model performance, we examined metrics like the log-likelihood, Akaike Information Criterion (AIC), and R-squared value. These can provide insights into the goodness-of-fit and power of the model.

We are making a key assumption that the predictors exhibit a linear relationship with the response variable. It is important to note that it may not fully capture complex interactions or nonlinear patterns in the data.



```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                      D   No. Observations:                51534
Model:                            GLM   Df Residuals:                    51530
Model Family:                Gaussian   Df Model:                            3
Link Function:               Identity   Scale:                          4754.6
Method:                          IRLS   Log-Likelihood:            -2.9129e+05
Date:                Tue, 03 Dec 2024   Deviance:                   2.4501e+08
Time:                        05:52:00   Pearson chi2:                 2.45e+08
No. Iterations:                     3   Pseudo R-squ. (CS):             0.2434
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        413.0801      0.939    439.946      0.000     411.240     414.920
Hour           4.9855      0.044    113.632      0.000       4.900       5.072
dotw          -5.3750      0.152    -35.374      0.000      -5.673      -5.077
month          1.2963      0.089     14.517      0.000       1.121       1.471
==============================================================================
```

Figure 9: Gaussian Frequentist GLM

Firstly, the Log-Likelihood measures how well the model explains the data, where values close to zero are preferred. Our model yielded a value of -291,290, which reflects the fit of the linear regression given the dataset. The AIC evaluates the trade-off between model complexity and goodness-of-fit, and penalizes models with too many parameters that might overfit. A smaller AIC value indicates a better balance. Our model's AIC value of 582,586, suggests that our model could use some improvement, by potentially adding more or less features. Lastly, r-squared provides an estimate of the proportion of variance in the response variable explained by the model, where values closer to one indicate a better fit. Our model output a value of 0.2434 is not terrible, but also not the best. It accounts for a fourth of the variance in electricity demand that is explained by time-based factors. While it is not particularly high, it highlights the influence of unmodeled variables, like weather or socioeconomic factors.

6.3 Bayesian Model

Now we shift to a Bayesian perspective, which might offer a complementary framework to understand our data. Bayesian regression allows the incorporation of prior knowledge or beliefs into the modeling process, providing a probabilistic interpretation of the results. Some key components of Bayesian regression are the prior distribution, likelihood, and posterior distribution. The prior distribution represents our beliefs or assumptions about the parameters before observing the data. We believe that the time of day (hour) is likely to have a strong, positive impact on electricity demand. The likelihood captures the information provided by the observed data. This is the same as our Frequentist approach, which is Gaussian, given the continuous nature of the electricity demand. The posterior combines the prior and likelihood using Bayes' theorem. This will give us parameter estimates, including credible intervals that quantify uncertainty.

The following model (Figure 10) was built using the Bambi library. We will model the following equation: $\text{identity} = \beta_0 + \beta_1 \times (\text{hour}) + \beta_2 \times (\text{month}) + \beta_3 \times (\text{dotw}) + \beta_4 \times (\sigma)$. We observed how well the predictors explain the observed variability in electricity demand.

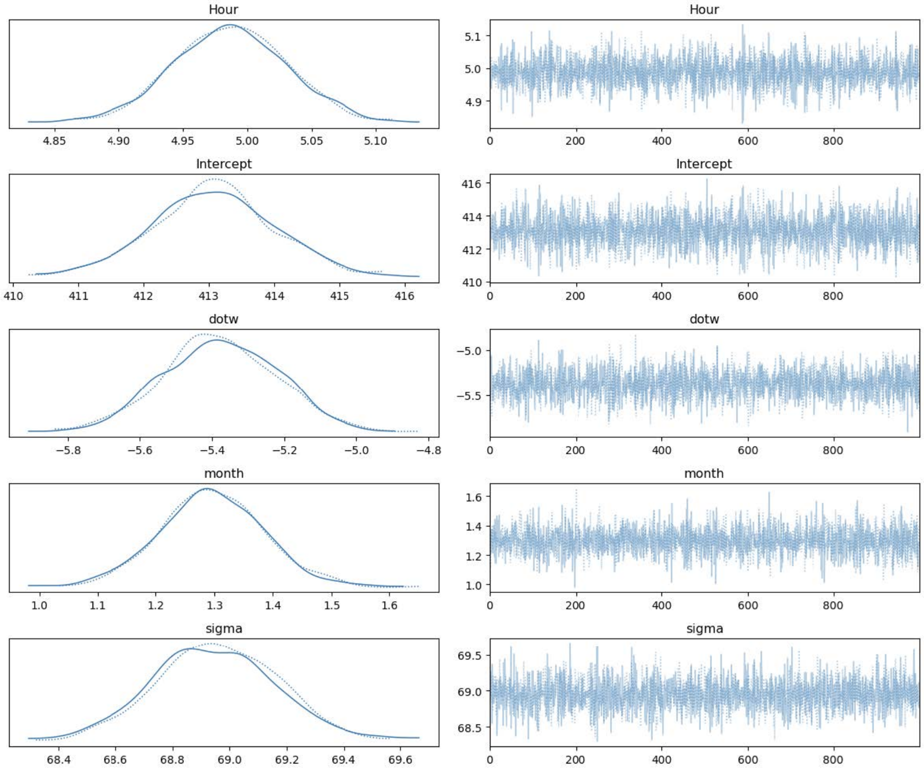


Figure 10: Bayesian Model

We are able to see what the Bayesian GLM outputs. The intercept coefficient is 413. The hour coefficient is right below 5. The month coefficient is about 1.3. The ‘dotw’ coefficient is about -5.4. And the standard deviation of the residuals (σ) is a little less than 69.

6.4 Bayesian Model: Posterior Predictive Check

We must now see how well our model captures the observed data. The Posterior Predictive Check (PPC) allows us to evaluate the model’s fit and ability to reproduce patterns in the observed data. The PPC involves generating sampled data based on the posterior distribution of the model parameters. We used the Arviz library to generate posterior predictive samples (Figure 11).

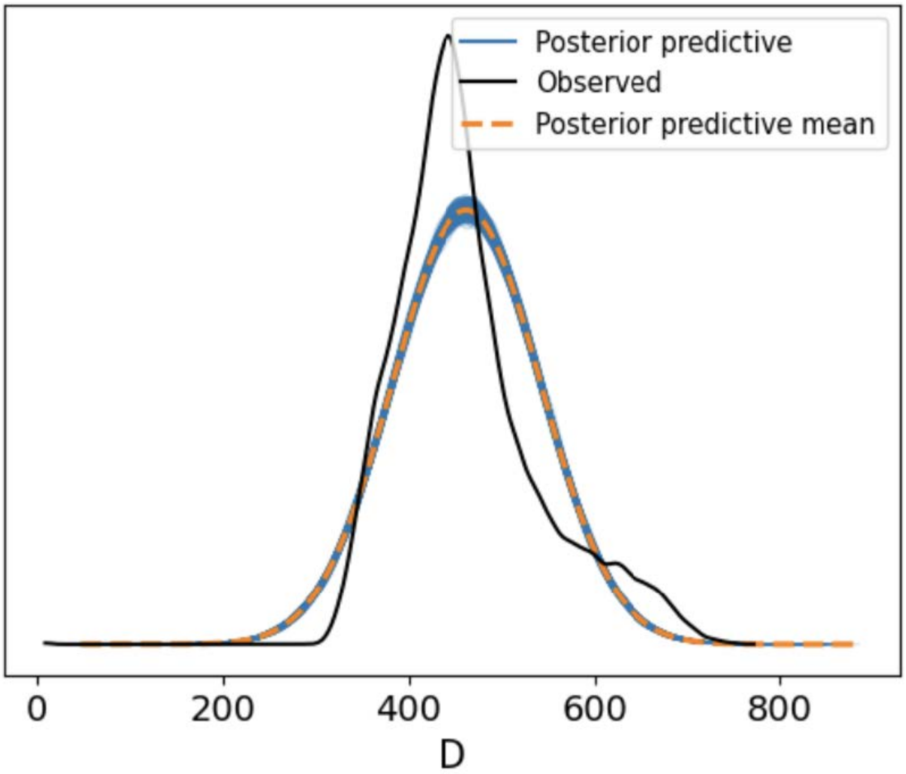


Figure 11: Posterior Predictive Check

The black line represents the actual distribution of electricity demand observed in the dataset. The blue line shows the distribution of simulated electricity demand values based on the posterior predictive samples. Although not perfect, the posterior distribution and observed data indicate that the model is capturing the overall shape and variability of electricity demand.

6.5 Bayesian Uncertainty

To quantify uncertainty, we can take a look at the credible intervals (Figure 12). We are answering the question that given the data, how uncertain is the unknown? These are the range of possible values that probably contains the unknown given the data.

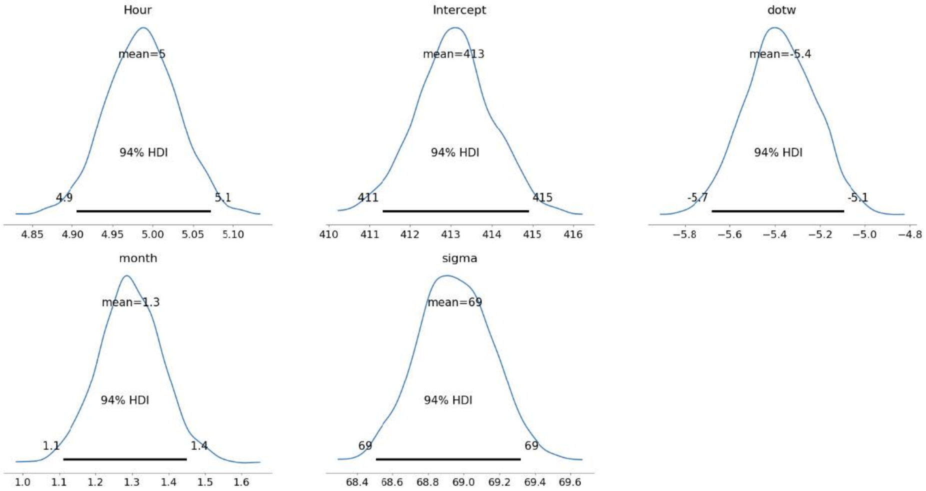


Figure 12: Credible Intervals

The X-axis represents the coefficient of each numerical variable and intercept in our Bayesian model. The Y-axis represents the probability density of the posterior density. Amongst all the graphs, the black line showcases the 94% Highest Density Interval (HDI). This shows the range of values containing 94% of the posterior probability. The roughly normal distributions show the effectiveness of the model. Now we look more closely into the parameter-specific values. Firstly in the 'hour' parameter, there is a 94% chance the coefficient will fall between 4.9 and 5.1, with a mean of 5. Meaning for every additional hour, the predicted demand increases by approximately 5 gigawatts. Next in the intercept parameter, there is a 94% chance the coefficient will fall between 411 and 415, with a mean of 413. Indicating that when all the predictors are zero, the baseline predicted demand is roughly 413 gigawatts per hour. Then in the day of the week parameter, there is a 94% chance the coefficient will be between -5.7 and -5.1, with a mean of -5.4. We can see with each additional day later in the week, the predicted demand decreases by about 5.4 gigawatts. Finally the month parameter, there is a 94% probability that the coefficient will be between 1.1 and 1.4, with a mean of 1.3. Meaning that for each additional month of the year, predicted demand will increase by about 1 gigawatt. It is important also to acknowledge the sigma (also referred to the error standard deviation), where the mean is 69. This represents the typical variation of demand not explained by the predictors.

Among all predictors, the hour parameter has the largest positive impact in electricity demand, with an increase of 5 gigawatts per hour. However, the other predictors show meaningful effects. Even small changes among the coefficients make a remarkable difference, given that one gigawatt is equivalent to 1,000 megawatts.

6.6 Non-Parametric Models

Now we shift our focus to non-parametric modeling. These are methods that make no assumptions about the underlying distribution of the data or parameters. They are well-suited to capturing complex and non-linear relationships in the data. We believe that decision trees and random forest regressors would best fit our predictions. Decision trees are intuitive and interpretable and random forests use a combination of decision trees to improve accuracy and reduce overfitting. To compare performance of the models amongst one another, we have chosen to calculate the root mean square error (RMSE) and R-squared. The RMSE tells us how far the model's predictions are from the true values, on average using the same units at the target variable, demand in our case. A lower RMSE indicates more accurate prediction and better model performance. R-Squared represents the proportion of variance in the target variable explained by the model. Values closer to 1 indicate that the model explains more variability in the data.

6.7 Decision Tree

Decision trees are one of the simpler non-parametric models that we have discussed in class. As we have seen in our initial EDA analysis, electricity demand is not entirely linear. We can capture trends effectively by splitting the data into smaller subsets based on feature values. For example, trees can split into categories like day and night or weekdays and weekends. In our analysis, we limited the maximum depth of the tree to 5, to make interpretability easier. We trained the model on 80% of the data, while leaving the remaining 20% for testing. To avoid any ambiguity in reproducing slightly different numbers each time we run our script, we set the 'random_state' to an arbitrary number of 42. Keep in mind that our visualization of the decision tree was truncated to maintain readability.

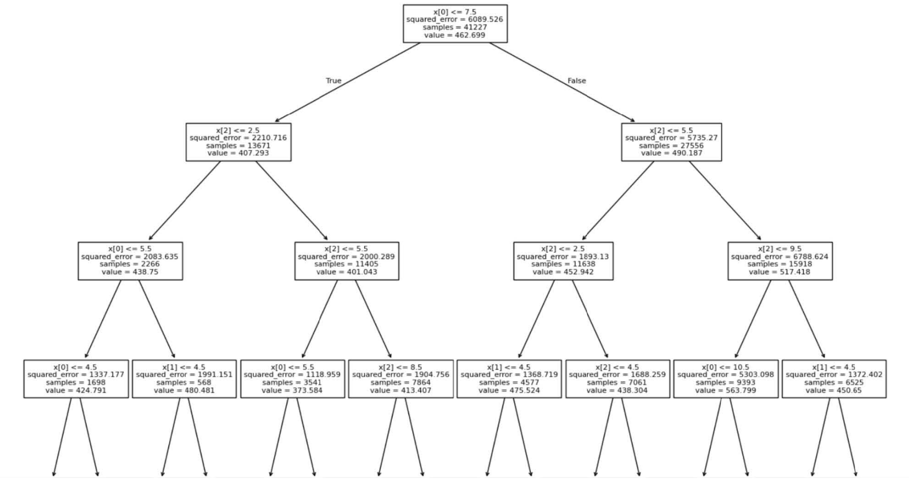


Figure 13: Decision Tree. [NOTE: In our code we implemented a depth of 5, however the visualization only shows 3. The rest of it is truncated for readability purposes.]

6.8 Random Forest Regression

While a simple decision tree does really well at visualizing the data at hand, we can improve our error metric performance. Random forests are a one-up from decision trees. They are an ensemble of decision trees that aggregate predictions. Each tree is trained on a randomly sampled subset of data with replacement. Once again, we will be using an 80/20 training and testing set and keeping the same number, 42, as our random state. Although this time, we are dropping any maximum depth to maximize our error metrics.

6.9 Non-Parametric Model Results

Random Forest Regression outperforms the Decision Tree model in terms of the root mean squared error (RMSE) and R-squared values. The RMSE are 29.97 and 40.11 respectively. The R-squared values are 0.85 and 0.73 respectively.

6.10 Discussion

When evaluating the effectiveness of the models, we compared the R-squared values for each one. The Frequentist GLM had an R-squared value of 0.2434, indicating that is explained roughly 24% of the variance in electricity demand. Both the decision tree and random forest regressors achieved R-squared values above 0.75, significantly outperforming the GLM. While, R-squared does not have a universal threshold for what constitutes a ‘good’ fit, values closer to 1 generally indicate better performance. Thus, R-squared can be used as a measure for absolute evaluation because it is bounded between 0 and 1. However, R-squared does not account for overfitting, so we must consider other metrics.

While there is not directly an R-squared metric for Bayesian modeling, we evaluated performance using a posterior predictive check (PPC). By visually inspecting similarity between empirical and simulated distributions, we found that the Bayesian model performs reasonably well. However, PPC is better suited for comparative evaluation rather than absolute judgment, since it lacks a clear numerical threshold.

Among the non-parametric models, random forest regressor outperformed the decision tree, with a higher RMSE and R-squared.

We would not recommend using the Frequentist GLM due to its poor goodness-of-fit metrics. The log-likelihood value was too far from zero and the deviance exceeded one million units. The model struggled to capture the complexity of the data.

We are fairly confident in applying the non-parametric models to future datasets, as they achieved significantly better performance. The models were trained on an extensive five years of historical data by the hour, providing robust insights. However, non-parametric models, particularly random forests, can be computationally expensive due to ensembles of decision trees. Additionally, decision trees alone are prone to overfitting, as they can achieve perfect accuracy on the training set.

The Frequentist model provided clear interpretations about how the predictors influence electricity demand. The positive coefficient in hour indicates electricity demand increases as the hour progresses. The negative coefficient suggests a decrease in electricity demand later in the week. And, the positive coefficient implies an increase in demand as the months progress.

The Bayesian model did not contain an interval that included zero. Thus, all the parameters included in the model have some linear correlation in regards to energy demand. There is a positive correlation among the hour of the day and month and a negative correlation for the day of the week.

The non-parametric models do not explicitly provide coefficients, however they still capture the importance and influence of features in hourly electricity demand variations.

The Frequentist GLM relies on assumptions about linearity. They might not hold our research question. The Bayesian GLM approach depends on prior choices, which can introduce bias if we are not careful enough. And non-parametric models are less interpretable than GLMs.

To enhance model performance, we can look beyond time-based features. Weather data could be beneficial by looking at temperatures, particularly when it is extremely hot or cold. Additional data about the region can capture underlying local patterns.

Overall, the uncertainty in our results are qualitatively high, primarily due to the continuous variable and inherent variability in electricity demand. There are many other confounding variables and predictors other than time that account for electricity demand. External factors like weather and policy changes are not captured.

Despite these challenges, non-parametric models that provide accurate predictions are well-suited for future applications.

7 Conclusion

In our first research question we explored how extreme heat events impact daily electricity demand in California. Through matching techniques, we found a significant positive causal effect in the increase of electricity demand on days exhibiting extreme heat.

In our second research question we aimed to predict hourly electricity demand using time-based features. By comparing generalized linear models (GLMs) and non-parametric models, we found that non-parametric models performed better, as they were able to capture more complex trends and outperformed in particular error measurements.

The electric demand dataset was originally categorized by region within the United States. However, it would have been helpful to have detailed information about which states were included in each region. The ambiguity of regional boundaries can vary from person to person. Although we were only looking at time, we should look for a deeper understanding of factors driving electricity beyond just the time-based features. One might ask an expert, *What are some external factors that influence electricity demand and how can we incorporate them into our predictive models?* Incorporating additional predictors, such as weather data and detailed regional information could enhance our analysis. Since we did not introduce these variables, we may have possibly introduced omitted variable bias. We also could have used more non-parametric models like gradient boosting or neural networks to capture even more complex and subtle patterns in the data. The results of the analysis are somewhat generalizable for electricity demand patterns for most of the United States, with the exception of Alaska and Hawaii. However, the findings can be interpreted as narrow due to exclusive reliance on time-based features.

Accurate electricity demand forecasts are essential for optimizing grid management, reducing energy waste, and preparing for peak energy demand and consumption, like heat waves as we have observed. Our call to action is for policymakers and energy companies to prioritize investments in data collection and analytics. These efforts would support energy efficiency initiatives and proper infrastructure maintenance and upgrades. Residents in regions prone to extreme weather would benefit by experiencing fewer outages. Upfront investment into these systems could improve long-term profitability by minimizing electricity shortages or surpluses within supply and demand.

References

- California Environmental Protection Agency, Office of Environmental Health Hazard Assessment. ”Extreme Heat Events.” Indicators of Climate Change in California, 2022. Accessed December 16, 2024.

- National Renewable Energy Laboratory. ”Extreme Weather Events.” Energy Analysis. *National Renewable Energy Laboratory.* Accessed December 13, 2024.

- Yukseltan, Ergun, et al. ”Hourly Electricity Demand Forecasting Using Fourier Analysis with Feedback.” *Energy Strategy Reviews*, 2020. https://doi.org/10.1016/j.esr.2020.100524