# Prediction of  Relationships Between  Energy Demand, Consumption, Air Temperature, and Extreme Weather Events

Nicholas Lemoff, Keita Tanabe, Areeya Tipyasothi, Erica Ying

# Data Overview

The temperature and humidity data for our first research question were collected as samples observed from weather observation stations within Alameda County. These data were collected on an hourly basis and aggregated to create daily averages for temperature and humidity over the whole county. Each row initially represented an observation from a particular station on a particular day but after our aggregations, each row simply represents the average for every day. This row representation may impact our results in that we may be missing important temperature variations across a day, which may cause us to misclassify whether a day is part of a heatwave or not. For example if there is a drastic nighttime drop in temperature, it may skew the daily average lower, making it seem as if it is not a heat-impacted time. We also may be missing temperature variations across locations that may have impacted our findings. Climate data sampled by the bay may be quite different from data sampled from a more in-land area, especially taking into account the various Bay Area microclimates. We ultimately decided that although there may be problems that occur due to missing data, since there are a large number of stations that sampled data, especially when compared with the amount of data that was considered to be possibly missing, it would not be significantly impactful on our findings. Thus, we disregarded the missing data rather than imputing. This was mostly a practical decision, as the total amount of data was extremely large, and any type of imprecise imputation of values may have adversely affected our findings more than simply discarding the relatively small amount of missing data.

With regards to the data used in our second research question, our preprocessing for climate data included aggregating across location and time to get to our desired granularity. We also joined together temperature and humidity data to generate an apparent temperature (what the temperature feels like), also known as the heat index via the heat index equation. If the heat index was above a threshold, defined as the 85th percentile of historical (between 1981 and 2010) temperatures in that region, then we classified it as 'heat-impacted'. If there were two or more heat-impacted days in a row, we considered those days to be part of a heatwave, in accordance with the EPA definition of a heatwave.

One concern with the accuracy of the data is the uniform treatment of nonuniform data. We weighted each data point collected at a given station equally (within our processing pipeline), and while the data is collected uniformly temporally by station, the stations are not distributed uniformly across the area of interest. There are disproportionately more stations close to population centers, increasing the effect of densely populated areas on our ultimate analysis. While this does not discredit our results, it is important to note that our findings are more applicable to areas that correspond to the denser clusters of data collection sites, and less to other areas.

Our emissions data was collected via census, however, at an extremely low granularity. The data fetched from the EIA database was from the emissions data for states by year, which we then paired with the total energy use by month in the state of California (also from the EIA database) to infer the amount of emissions per month. Due to the relatively high granularity of our data, we did not see any issues with missing data, and if there was missing data, it would have been handled by the EIA database before we accessed it.

The assumptions made in this were that the proportion of emissions per month was the same as the proportion of total energy used in the year in a given month. This could be a faulty assumption in a few ways, including that the proportion of energy generated by methods that generate emissions (ie natural gas, oil, or coal) and energy generated without emissions (ie. solar, wind, etc) in a given month may be variable based on the conditions of the given month (amount of sun, wind, etc.), and that the energy consumption data is only representative of the electricity consumption of a given period. This means that other sources of emissions such as gasoline used in transportation are not included in our calculation of emission proportions according to month. Effectively, this assumes that all emissions are proportional to electricity consumption for a given month. One significant issue that may occur is the underestimation of emissions in the cold months of the year, as our electricity demand does not capture the use of natural gas by end consumers in home heating solutions, which would account for a nontrivial amount of energy consumption and greenhouse gas emissions.

We made this assumption partially due to necessity, being that there was not a way for us to use the data available in the database to consider all sources of emissions in our proportioning of the emissions. We also felt that the proportionality of energy consumption was fair on two bases, being that a large portion of transportation would be a relatively constant source of emissions, which could be captured in our models using a constant bias feature, and otherwise, it would generally follow the trends of energy consumption due to the similar use cases.

## Research Questions

### 1. How can we predict impacts to the energy grid during extreme weather events?

By attempting to predict impacts to the energy grid during extreme weather events (heat waves, specifically for our research question), we may be able to make recommendations about responsible energy usage to consumers in anticipation of these weather events. We may also be able to recommend methods for proper resource allocation to balancing authorities. By anticipating when/where we can reduce demand and increase distribution, we could reduce stress on the energy grid, which would reduce the possibility of grid failures, unplanned blackouts, and other cascading effects.

Generalized linear models and non-parametric methods are a good fit for answering this question because they are suited for modeling non-linear relationships and both categorical and numerical outcomes. With these methods, we would be able to predict whether something is “impacted” (defined as the 90th percentile of energy demand for our purposes) as well as the actual value of the demand itself. It is important to keep in mind the limitations of these methods however: GLMs carry heavy assumptions about the distribution of our data and non-parametric methods carry the risk of overfitting. Thus it is important to thoughtfully select our GLMs and consider hyperparameters that will reduce overfitting for our non-parametric methods.

### 2. What is the causal effect between an increased demand for electricity and average air temperature in California?

Measuring the causal effect between demand for electricity and air temperature is an important area of research, as it can help us inform decisions that can impact the environment and our quality of life. By establishing a connection between a variable which we control (energy demand), and one that has immediate consequences on our lives (air temperature), we would hope to encourage decisions to be made to help control energy demand, or transform the sources of the energy to reduce emissions. The relationship is well understood in one direction, being that average air temperature drives demand for electricity due to use cases such as climate control and cooling of equipment. This is immediately clear from our research into prior work and our exploratory data analysis. The other direction has several layers of separation, however, and thus is harder to immediately draw causal conclusions for.

Using causal inference here is important to answering our question, as we are trying to establish a link that is stronger than a correlation that we already know exists. Since we already know there is a strong link between the two variables, using causal inference here will help us to draw stronger conclusions. Our method may be limited by the amount and granularity of data which we are able to procure, and the understood causal link in the other direction may interfere with our analysis. If we are unable to isolate the impact in the direction that we want, it may be hard to draw strong conclusions.

## Prior Work

https://iopscience.iop.org/article/10.1088/1748-9326/11/11/114008/ampdf
Matthew Bartos et al 2016 Environ. Res. Lett. 11 114008
https://downloads.globalchange.gov/sap/sap4-5/sap4-5-final-all.pdf (pg 7-29)
Scott, M. J. and Y. J. Huang, 2007: Effects of climate change on energy use in the United States in Effects of Climate Change on Energy Production and Use in the United States. A Report by the U.S. Climate Change Science Program and the subcommittee on Global Change Research. Washington, DC.

Energy generation and consumption are both deeply connected with temperature. We see that there is a wealth of research and knowledge on the causal effect of temperature on energy generation and consumption, which makes sense. There is a clear connection where our energy infrastructure runs less efficiently when ambient temperature is higher, and energy demand naturally rises when temperature is higher due to common uses such as climate control.

One of the main indicators for efficiency of our energy generation is how much energy generated has to be allocated to cooling the infrastructure to make sure that it does not break down. Our energy generation technologies are not perfect, so no matter what methods we use, there will be excess energy that escapes in the form of (primarily) heat. This is a problem for our infrastructure, as an excess of this heat can cause less efficient energy transfer, or in extreme examples, meltdowns. This means that the production capacity of our  power plants is dependent on the temperature of the system, which is in turn affected by the ambient temperature caused by the weather.

However, energy production is not the only concern to our infrastructure, energy delivery is just as impacted by rising temperatures. The study by Matthew Bartos, *Impacts of rising air temperatures on electric transmission ampacity and peak per-capita electricity load in the United States*,  tackles this concern. His findings were that within the next 30 years, we would find an increase in energy consumption peak due to temperature increase of 4.2-15% on average. The study also predicted a 1.9-5.8% decrease in summertime transmission capacity. Ultimately, while the study was not concerned with efficiency, it outlines a key reason to believe that efficiency would be impacted. Higher operating temperatures, as well as degraded energy delivery cables would cause more energy to be lost during delivery, which would in turn negatively affect the efficiency of the system.

In addition, the study by Bartos shows a link between temperature and energy demand. The greenhouse effect links air pollution caused by emissions to increases in temperatures, and the link between energy demand and emissions. This gives an intuitive causal link between energy demand and temperature, which we are attempting to study.

The report by the US Climate Change Science Program and Subcommittee on Climate Change Research, *Effects of Climate Change on Energy Production and Use in the United States*, explains the connections between climate change and energy use and production, particularly focusing on the direction of climate change affecting the other two, rather than the other direction. One consideration that this study brought to our attention was that of end use of natural gas for heating, which we had failed to consider before, however, as pointed out by this report, as global temperatures rise, the use of heating consumes less energy, and the large energy pulls of colder regions and times of year are actually more related to lighting and equipment rather than heating, thus, rather than natural gas dominating energy consumption, it would still be electricity.

The research in this report also supports our assumption that emissions would be proportional to energy consumption, as the main source of energy demand, air conditioning, is expected to grow in the same way in sectors such as transportation, which are not captured in our data, as in electricity, which is captured in our data. However, despite this, the authors were also unable to find conclusive sources of data on energy consumption by the transportation sector with regards to specific use cases such as air conditioning, and thus, their findings are not entirely conclusive.

While there is a section on energy production and distribution in the report, it is primarily concerned with the availability of resources rather than the efficiency of our electricity generation methods in a warming world, so it is not directly applicable to our area of research.

# EDA

Figure 1.a: Daily electricity demand over 2019 with impact indicator

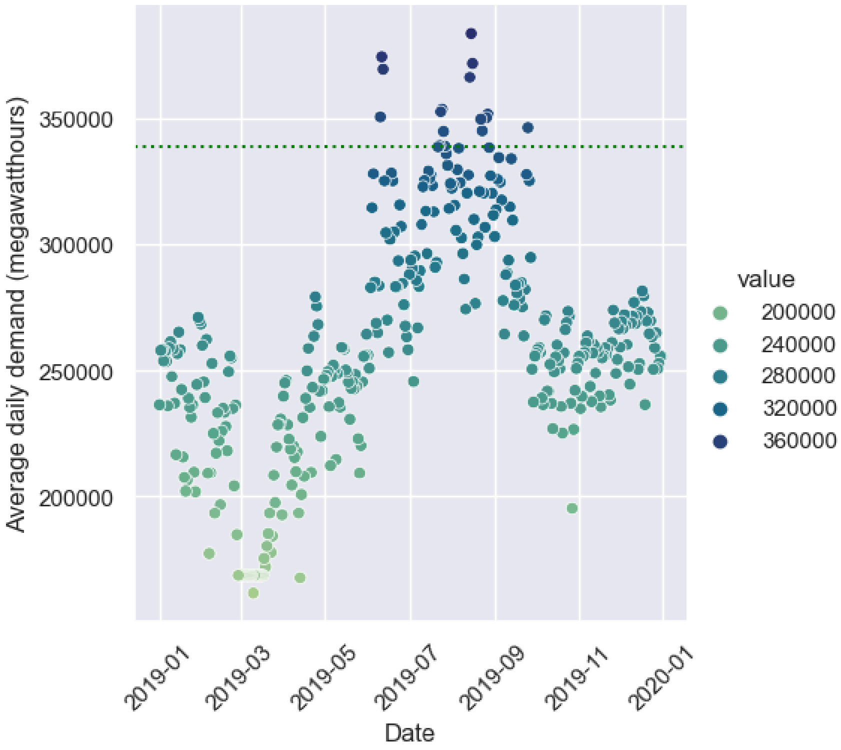


Figure 1.b: Daily 2019 heat index with heatwave indicator

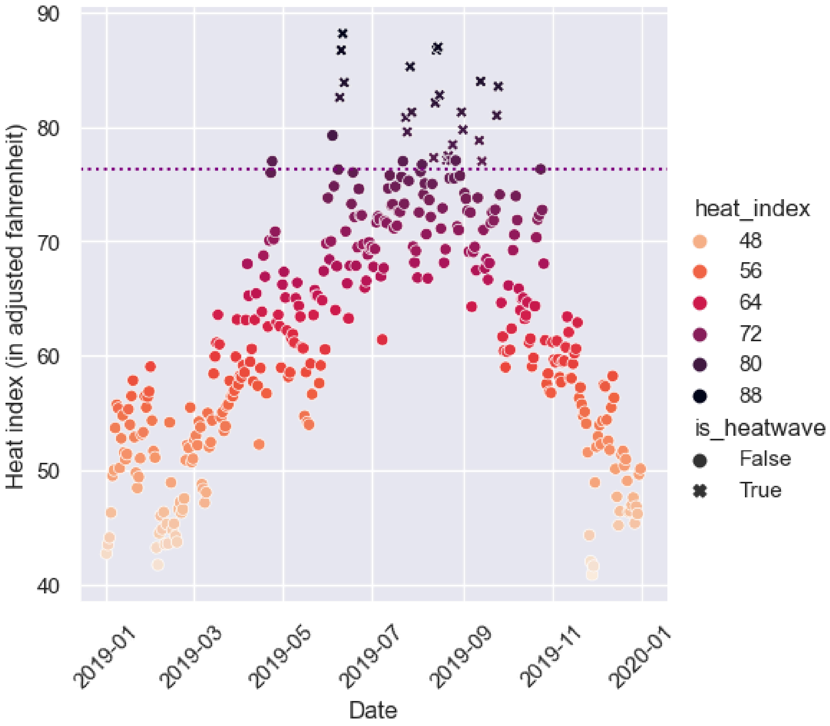


Figure 1.c: Relationship between electricity demand and heat_index

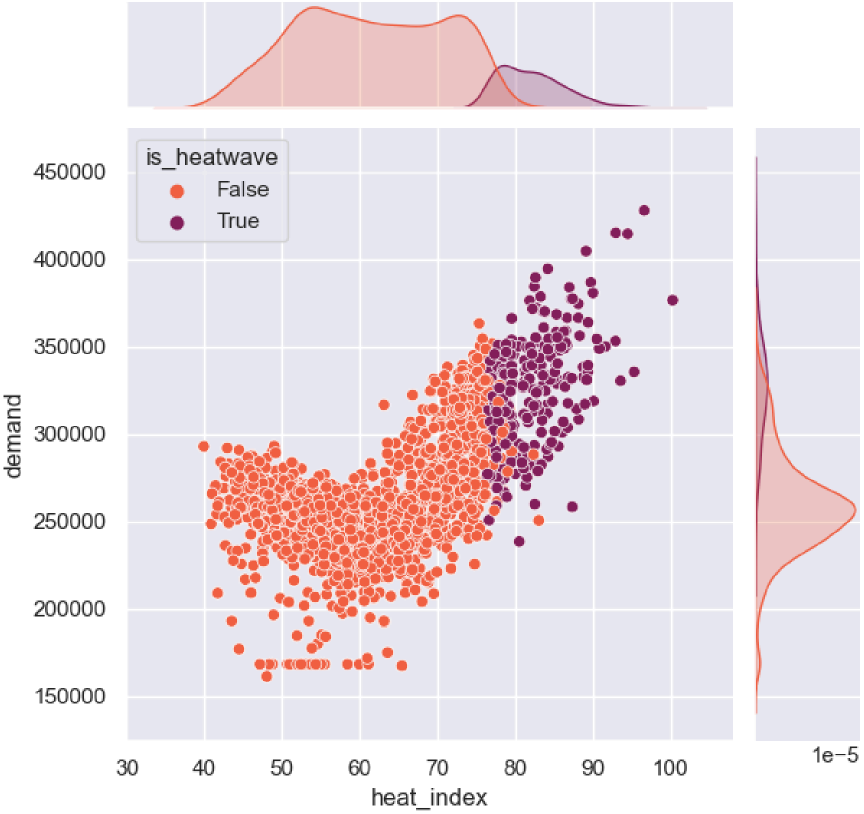


Figure 1.d: Emissions (thousands of tons of CO2) by month in 2023

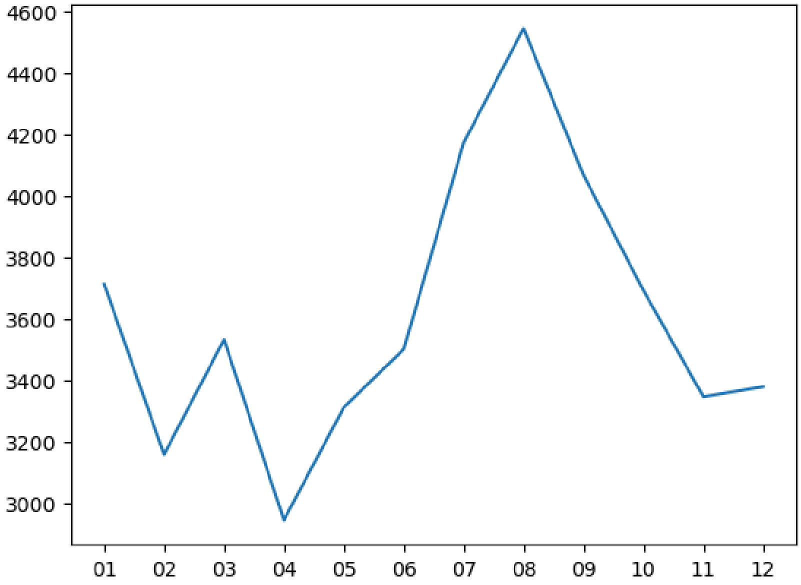


Figure 1.e: Temperature (C) in 2023

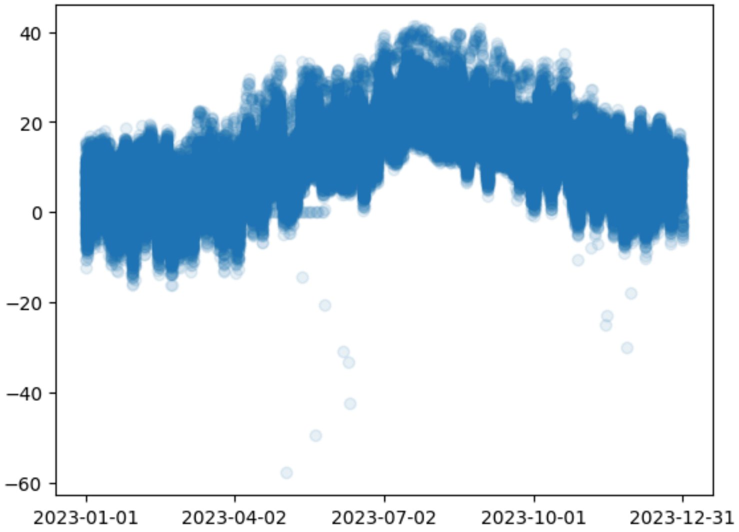


Figure 1.f: Average state-wide temperature (blue) and energy consumption (red) in 2023

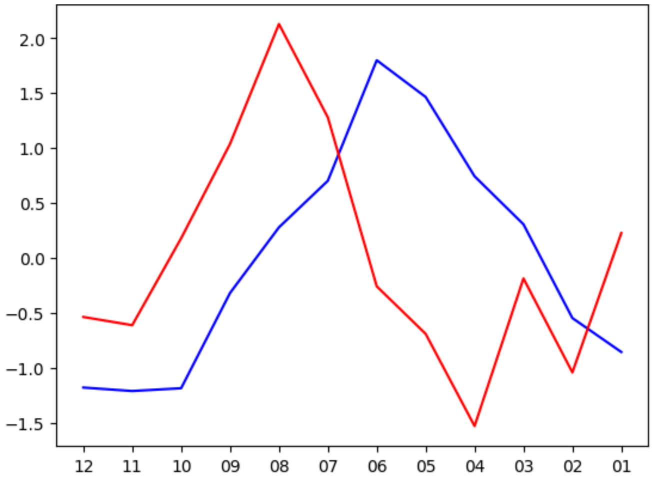


Figure 1.g: Average state-wide air temperature (blue) and industrial energy consumption (red) in 2023

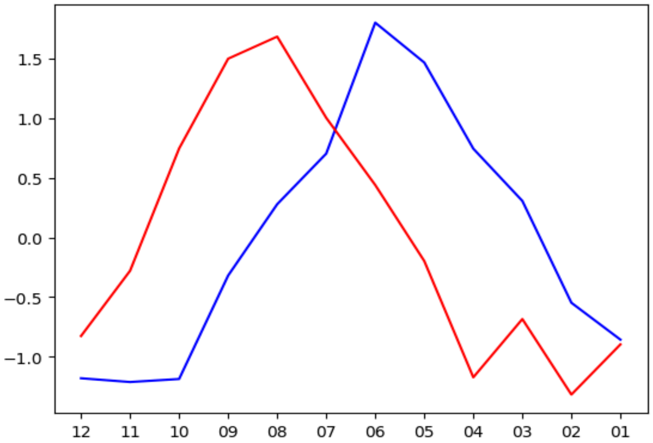


Figure 1.h: Average state-wide air temperature (blue) and transportation energy consumption (red) in 2023

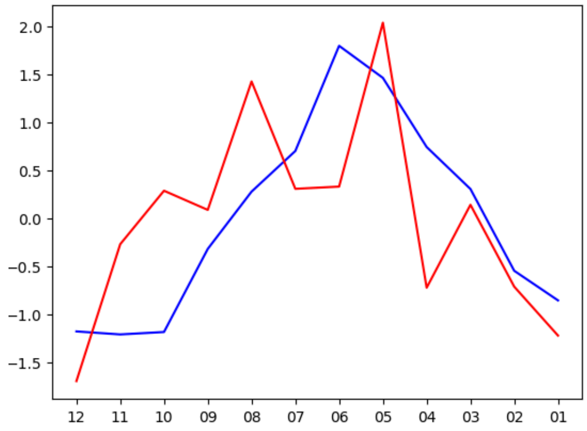


Figure 1.a shows consumer demand for electricity in megawatt-hours over 2019 aggregated by day while figure 1.b shows the daily heat index, also known as apparent temperature (what it “feels like”), over the same period of time. By comparing the two figures, we see that there does seem to be some positive correlation between the demand for electricity and heat index. This is further supported by a joint plot of the two variables in figure 1.c, which seems to indicate that with some transformation, a GLM may be quite effective in predicting demand from heat index. These visualizations give us a preliminary look at how related demand and heat are, which gives us more information to motivate our model selection.

A similar, seemingly positively correlated relationship can also be seen in figures 1.d and 1.e, where emissions appear to peak around the same time of year as temperatures (around late summer). As our second research question concerns the connection and potential causal relationship between temperature, consumption and emissions, this correlation is a promising look at whether there is a relationship between these factors.

The last three plots (figures 1.f, 1.g, 1.h) provide more insight into this relationship, with 1.g and 1.h consumption disaggregated by two types: industrial (manufacturing, mining, processing…) and transportation (motor vehicles, aircrafts, maritime systems…). The observed correlation between averages still appears to hold when divided up by these sectors, especially that of the industrial sector which we know from prior research, accounts for a majority of energy consumption. Especially notable is the shift in the peak value, with the maximum temperature for the year immediately following the maximum industrial consumption of the year. While certainly not indicative of causation, it is a possible look at how temperatures may be reacting to consumption and/or associated variables.

## Question 1: Prediction with GLMs and Nonparametric Methods

### Methods

**Goal**: To predict whether daily energy demand in Alameda County will be impacted (defined by exceeding the 90th percentile of historical energy demand) by heatwaves.

**Features**:
- tmax, tmin, avg_temp: maximum, minimum, average daily temperature.
- humidity: daily average humidity.
- heat_index: apparent temperature calculated by NOAA’s formula.
- is_heatwave: binary indicator of heatwave days based on EPA’s definition.

These features are directly related to environmental factors that influence energy usage. For example, high temperatures and increased humidity can potentially lead to elevated cooling demands, thus impacting energy usage.

**GLM Approach:**
We use a logistic regression model to predict the binary target *is_impacted*, which indicates whether the energy demand exceeds the 90th percentile threshold. Linear regression model is commonly used in classification scenarios because its link function could effectively transform a linear combination of input features into probabilities constrained between 0 and 1.

Link_function: logistic link function with Bernoulli likelihood.
Assumptions:
1. The model assumes that daily energy demand is independent of other times, which is consistent with the context of daily forecasting, but temporal autocorrelation is not considered.
2. Complex interactions between features are not accounted for in this model.

**Non-parametric Approach:**
1. K-Nearest Neighbors (KNN): Use KNN to identify energy demand spikes by looking at proximity in features like temperature and humidity. For example, days with similar temperature patterns and heat index profiles tend to demand comparable energy.
2. Decision Tree: It can capture complex, non-linear relationships between features, and is thus useful in cases where the relationship is nearly deterministic.

**Model Evaluation**:
- Split the dataset into training and validation sets to test the model.
- Use metrics like precision, recall, f1 score and AUC to evaluate the different models.

## Results

| | **Logistic Regression** | **KNN** | **Decision Tree** |
|---|---|---|---|
| Accuracy | 92% | 92% | 91% |
| Precision | 75% | 72% | 75% |
| Recall | 44% | 56% | 37% |
| F1 score | 0.55 | 0.63 | 0.49 |

| | | | |
|---|---|---|---|
| AUC | 0.954 | 0.926 | 0.937 |

While we had high accuracies across our three models, our precision and recall were much lower. This means that our models were able to mostly accurately classify whether a day was energy impacted (*is_impacted*) based on our selected features, but they were not as able to say whether a predicted impacted day was actually an impacted day, nor were they able to precisely predict the actual amount of impacted days out of all the actually impacted days. Unfortunately this indicates some issues with our models. Considering that most days will not be impacted days, the models could have easily achieved high accuracy by predicting that the day is not impacted. This approach would explain the disparity between our metrics, and is especially illustrated by the results of our decision tree model, as decision trees are much more prone to overfitting, which has resulted in the biggest jump between accuracy (91 percent) and recall (37 percent) out of all our models.

## Discussion

**Model Performance:**

- Logistic regression: performed the best overall, achieving the highest AUC and a balanced accuracy of 92%. However, the relatively low recall highlights a limitation in capturing all true impacted cases.
- KNN: slightly less robust in AUC, but achieved the highest recall among the models. However, its reliance on proximity in the feature space means its performance may be affected by noises in the dataset.
- Decision Tree: the recall is particularly low, indicating lower ability in identifying true impacted days. The hierarchical nature of the decision tree makes it susceptible to overfitting, which may prevent it from generalizing to the test set.

**Result Interpretations:**

The logistic regression model coefficients indicate that higher heat_index and humidity are strongly associated with higher probabilities of impacted energy demand, which aligns with our initial hypothesis. For KNN and decision tree, the interpretation is less explicit, but the high recall of the KNN suggests that it captures subtle patterns missed by the logistic regression, which may be due to the interaction within the features.

**Limitation of models:**

- The logistic regression model assumes linearity, which may oversimplify complex relationships.
- The KNN method requires more computational power, which may be expensive for analyzing large datasets.
- The decision tree model may cause overfitting, especially when the size of the dataset is limited.

**Future work**:

We can incorporate more features such as real-time energy prices, temporal factors like holidays, and infrastructure metrics like air conditioning coverage. We can also explore other modeling approaches like bagging and boosting, or RNN, which can capture more complex relationships and temporal dependencies in the data. We can also use oversampling or undersampling to address the sample imbalance issue. In addition, we could also extend the study to multiple regions with different climate conditions and apply cross-validation to improve the generalizability and stability of the model.

# Question 2: Causal Inference

## Methods

**Treatment Variables:**

- Definition: Electricity demand in megawatt-hours.
- Hypothesis: Electricity demand helps drive carbon emissions because of increased generation requirements.
- Mechanism: Greater electricity demand often leads to additional power generation from carbon-intensive sources. This relationship is influenced by temperature trends (Figure 1.e), which impact electricity demand.

**Outcome Variables:**

- Definition: Carbon emissions, measured in metric tons.
- Goal: Try to capture the impact on the environment associated with changes in energy demand.
- Objective: Establish a causal link between demand patterns and emissions.

**Confounders:** In this case, confounders are going to be variables that impact both electricity demand and carbon emissions independently.

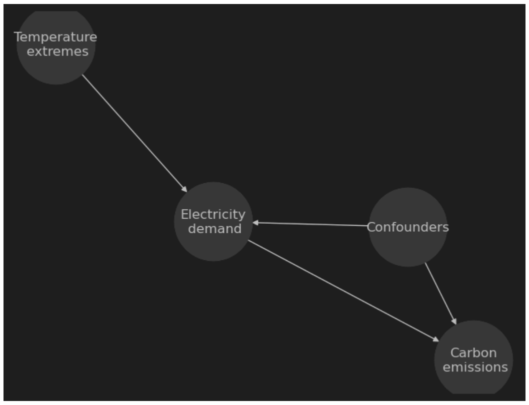
- Season: There is a seasonal impact on the air temperature (Figure 1.e). It is understood that there is a similar impact on energy demand and emissions. (Figure 1.d)
- Economic activity: Having increased industrial activity increases electricity usage and also contributes to emissions. (Figures 1.f, 1.g, 1.h)

**Unconfoundedness Assumption:** For our unconfoundedness assumption to hold, the instrumental variable in this case, temperature extremes, would have to influence the outcome, being air temperature and emissions, only through our treatment variable of electricity demand. This argument is based on both physical and social responses to change in temperature. To give an example, if there is a heat wave, cooling demands will go up, which raises electric consumption, ultimately impacting carbon emissions. Since the temperature doesn’t directly affect emissions outside of its impact on energy usage, this assumption is plausible.

**Adjustment for Confounders:** We decided to use the instrumental variable methods to address confounding. In this case, we used temperature extremes because it is a strong predictor of energy demand. Regression models were centered around using temperature and non-linear transformations like square root terms to model this complex relationship. By attempting to isolate variation in energy demand that can be attributed to temperature, the instrumental variable approach mitigates bias from confounding variables.

**Colliders:** The main collider we’d have to consider in the dataset is economic activity. For example, if there is an extreme heat wave in the Central Valley, farmers might have to work less hours since it’s too hot. This shows that economic activity is influenced by temperature, and drives both energy demand and emissions. To prevent these collisions, our analysis excluded economic activity from direct adjustments and relied on the instrument to account for indirect effects.

**Causal DAG**



## Results

**Summary and Interpretation:**

- First-stage regression: Showed a significant positive relationship between the temperature and electricity demand. In this case, higher temperatures led to increased energy usage, most likely from people trying to power their air conditioner units. The temperature coefficient was statistically significant.
- Second-stage regression: Looking at the impact of predicted energy demand on carbon emissions and the potential effect on air temperature resulted with a coefficient of about 0.1811. The implications of this is that for each unit increase in energy demand (megawatt-hours), $CO_2$ emission increased by 0.1811 metric tons. With that said, the $R^2 = 0.681$ which indicates that even though the model explains a decent amount of the variance, there is still a good amount of variability that isn't accounted for.

**Effect Magnitude:** Although the magnitude is statistically significant, it suggests a somewhat modest increase in emissions per unit of electricity demand. This makes sense since California gets energy from a variety of different methods, and renewables may offset the carbon intensity of generating electricity.

| Regression | Independent Variable | Dependent Variable | Coefficient | $R^2$ | p-value |
|---|---|---|---|---|---|
| First-Stage | Temperature | Electricity Demand | 47.24 | 0.695 | 0.012 |
| Second-Stage | Predicted Electricity | Carbon Emissions | 0.1811 | 0.681 | 0.001 |

| | | | | | |
|---|---|---|---|---|---|
| | Demand | | | | |

### Discussion

**Limitations:**

- Extrapolation Assumptions: Our analysis assumes uniform efficiency in generating electricity throughout the year. This doesn’t account for how the breakdown of how energy is made may change throughout the year (e.g., more solar energy is generated in the summer).
- Spatial and Temporal Variability: We did not account for urban rural divide or the geographic diversity that characterize California.
- Data Granularity: Our data was aggregated monthly which doesn’t account for more short-term fluctuations and potentially decreases sensitivity to causality.

**Additional Data Needs:** For our experiment there are a few different datasets that we weren’t able to source that would have improved our results. This includes a detailed monthly breakdown on renewable vs. non-renewable energy production, a better measure of industrial activity, and a better geospatial breakdown of emissions.

**Causal Confidence:** The positive relationship we observed between electricity demand and emissions supports a causal interpretation. However, full confidence in these conclusions is difficult due to limitations with our model and data. It is possible that there are alternative explanations like changes in the composition of energy demand shifting from residential to commercial or vice-versa.

**Recommendations and Next Steps:** Future work on this should try to get higher-resolution and more time-specific data. This would aid in capturing short-term variations in both energy demand and emissions. Delving further into non-linear modeling techniques would also help with modeling more complex interactions between variables. Another thing to consider would be analyzing regions within California which could help us gain insight into spatial heterogeneity in demand and emissions.

## Conclusion

### Outcomes Summary

This study aimed to predict impacts to the energy grid during times of extreme weather and establish a causal relationship between energy demand and air temperature in California. The key findings include:

**Prediction Models:**

- Using logistic regression achieved the best precision/recall score for identifying days of energy grid stress from heat waves with an accuracy of 92% and an AUC of 0.954. For the amount of data available, logistic regression proved to be the best choice on all ends, as it is also the simplest to understand intuitively.

- Using KNN showed higher recall, which is good for capturing more subtle patterns but required more computing power. It is also harder to interpret the results of the model, as the operations are very complex and do not have simple interpretations.
- Using decision trees did not work for the data well, likely because the model was much too robust for the data available, leading to overfitting, low precision on sets other than the training set, and without the interpretability of a GLM.

**Causal Analysis:**

- The instrumental variable approach that we used demonstrated a modest but still statistically significant impact of electricity demand on carbon emissions. It had a causal coefficient of 0.1811 metric tons of $CO_2$ per megawatt-hour.
- Temperature extremes proved to be a robust instrument for modeling electricity demand.

## Critical Evaluation

Many of our limitations were due to data that were mismatched in granularity or lacking in extra key information. For example, our climate data was taken from weather stations which we aggregated across our area of interest to gain an average over the location, while our emissions and other energy data were aggregated over a balancing authority region. Furthermore we did not take into account variations in extreme weather events (heatwave versus a hurricane) nor did we take into account what kind of emissions were being released and from what source. We also limited much of our analysis to a period of a year and were not able to do an analysis over time to verify that our results would be the same between years, nor were we able to compare whether an increase of emissions per year has led to a multi-year increase in air temperature.

Due to the limitations above, we do not think our conclusions and models are as robust as they could be. One way that we could fix this is by including a larger range of time to train and validate our models on, as to compare whether variations over each year could be responsible for changes in air temperature or energy demand. Thus we do not believe our results to be generalizable to any other year or location.

We could possibly ask a climate expert whether they are aware of any known causal pathways for determining increases in air temperature, aside from the obvious time of year. By understanding more about already observed causal factors, we may be able to further validate previously done studies to confirm that these factors are indeed causing changes in air temperature, rather than using our instrumental variable of temperature extremes. This extra domain knowledge would be quite helpful in informing our causal analysis and potentially allow us to avoid considering the confounding variables that may have affected our research.

## Recommendations

Moving forward, while our findings have significance, there were limitations in our equipment and resources that limited our ability to conduct our research in the way that we wanted to. We would recommend that similar studies be conducted on larger time periods so that the confounder of seasonal shift in temperature can better be accounted for, and that the data be considered on a more granular level, possibly on a smaller area of land. These were not possible for us, as the only energy data that our machines could process was that of the very generalized monthly and yearly consumption numbers over the whole state. This led us to use temperature data from the whole state, which greatly impacted the specificity of our data and the ability of our models to produce meaningful results. In the future, increasing the granularity of one set or decreasing the granularity of another would help greatly to better models.

With all of this said, we have observed causal links between electricity demand and CO2 emissions, which shows that the relationship does not go only in one direction. This is important because it shows us that we need to focus on ways to break the positive feedback loop that is being created. We recommend taking action to decrease the dependence on climate control. This can take the form of restricting the use of climate control in businesses, imposing stricter restrictions on insulation, and investing in more energy efficient technologies to meet the needs of the population.

We also observed the use of heat index in the prediction of energy demand, so we recommend that this indicator be used in decision making going forward. Observed heat indices should be treated as serious occurrences, and the state should issue notices to mitigate energy consumption by those who do not need it during those times. It would also be beneficial to invest into studying ways to predict heat index with high accuracy, so that we have more time to prepare for extreme weather events.