2023

WeatherWatts Project Report

EDWARD OH | ISLAM ORABI | SHANE SARABDIAL

Contents

Introduction

In the U.S., the energy demand and its price have surged in the last few years due to political, environmental, and geological factors. Our motivation for choosing this topic is due to extreme weather cases that have become more frequent in the last few years. This project investigates key features that may impact energy demand in New York, California, Texas, and Florida. We analyzed statewide energy demand using hourly weather information from New York City, Los Angeles, Austin, and Tampa.

Additionally, we explored each state's population size and area as possible factors that may impact the energy demand. We constructed a predictive model to forecast energy demand for three days in the future based on previous hourly energy demand, weather information, and energy stock prices. This capstone will answer the following questions:

1. How much of an effect does the weather have on energy demand?
2. Does the cost of energy reduce demand?
3. When does energy demand peak, and when does it subside?
4. What does the energy demand look like leading up to, during, and after a natural weather event?
5. Does population affect energy demand?
6. Is there a correlation between state area and energy demand?

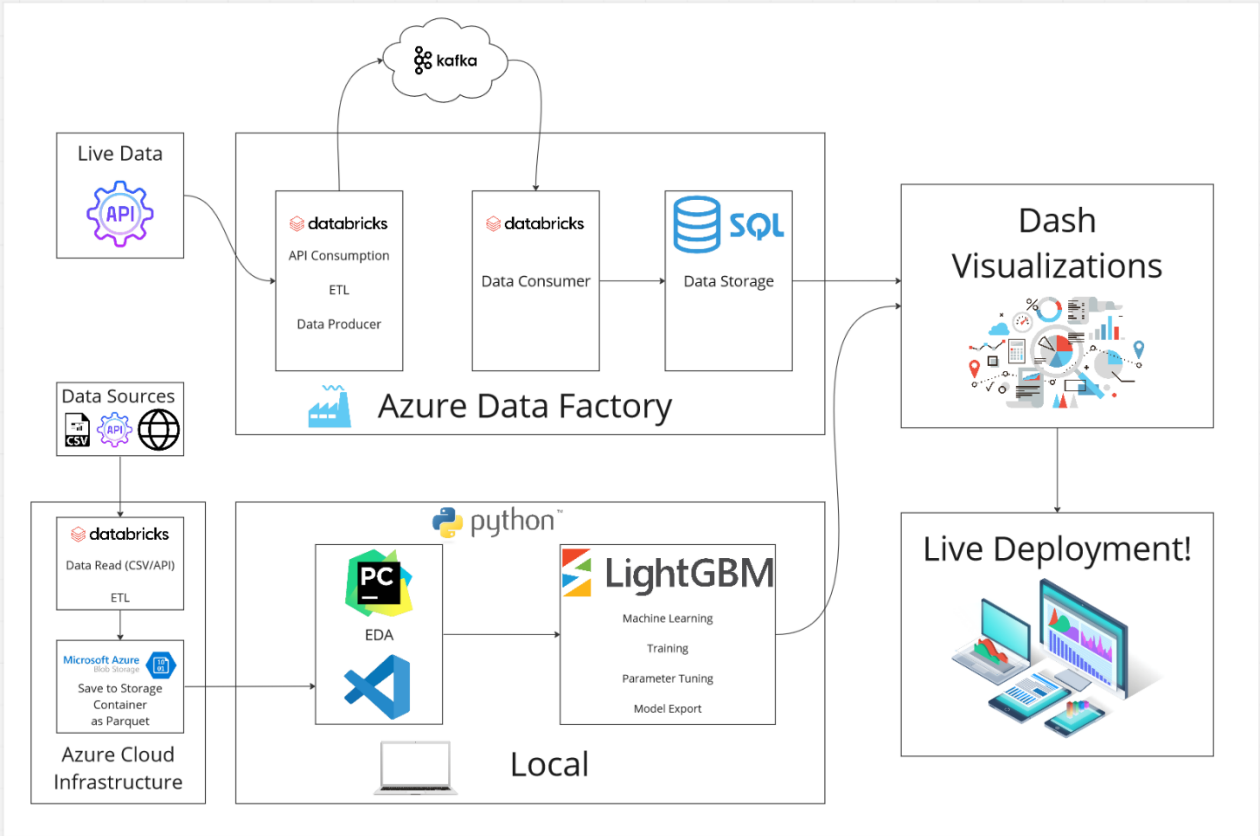
Data Sources and ETL

The primary source of our data comes from the U.S. Energy Information Administration (EIA). We accessed data through the EIA's API. We got weather data from Visual Crossing, a website that collects historical weather data. We accessed historical and forecasted data through Visual Crossings' API. XLE is an index that tracks energy companies. We used Yahoo Finance to get the stock price for XLE data through a CSV and an API. Our last main data set was a list of holidays scraped from Timeanddate.com using Beautiful Soup. The energy demand and weather were hourly data, while XLE was daily.

We extracted two supplementary datasets from the United States Census Bureau and The Fact File for yearly state population size and state area, respectively. An energy generation dataset from the think tank's website, Ember, was extracted to investigate further the correlation between state area and energy demand for all states. The state population size dataset from the United States Census Bureau and the energy generation dataset from Ember were extracted as CSV and loaded into separate data frames for analysis. The state area data from The Fact File was web-scraped using Beautiful Soup and converted into a data frame for analysis.

Pipeline

Static data was pulled from 4 different sources utilizing API, CSV downloads, and web scraping. We cleaned and saved the data in our Azure cloud storage container. We used that clean data to do exploratory data analysis and create our machine-learning model with LightGBM. Every 24 hours, live weather data, XLE price data, and EIA energy demand is pulled using an API to be cleaned, produced, and consumed in data bricks and Kafka. This process is automated using Azure Data Factory. The ingested data is stored in an SQL database to be inputted into our machine-learning model. Visualizations are created using Dash and publicly displayed on our website.



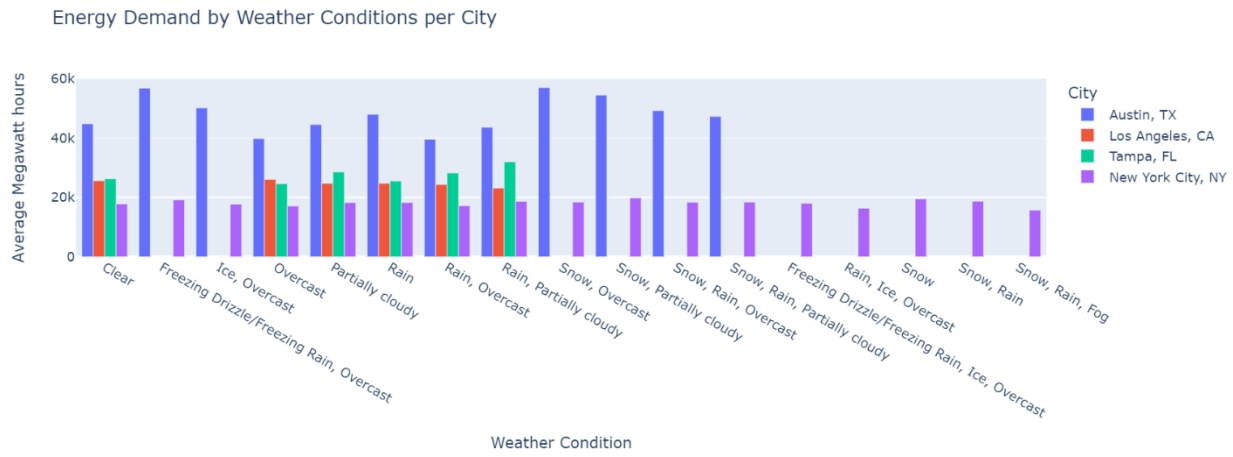
EDA

Energy Demand and Weather Conditions

Our investigation began with a look at the weather conditions in each city and comparing the average energy demand in their respective states. The city's weather conditions give an overview of each state's general weather conditions.

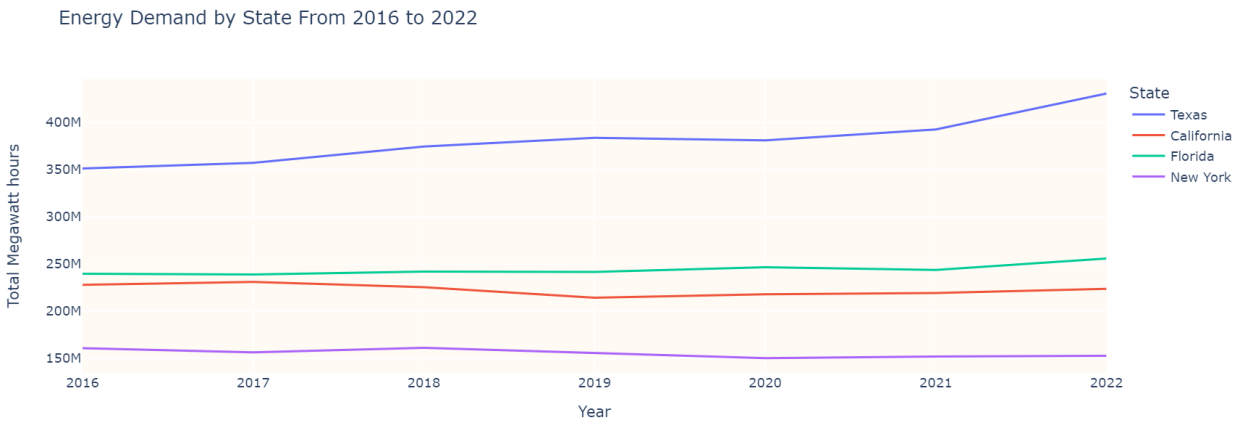
This bar chart shows the hourly average energy demand in megawatt hours in Texas, California, Florida, and New York based on weather conditions in Austin, Los Angeles, Tampa, and New York City. In Austin, when there is snow and overcast, the average energy demand in Texas is high, while a weather condition of rain and overcast yields a low energy demand. In Los

Angeles, when the weather is overcast, the energy demand is highest in California and lowest when the weather is rainy and partially cloudy. As for Tampa, when the weather is rainy and partly cloudy, the energy demand in Florida is highest, and when the weather is overcast, the energy demand is lowest. For New York City, the energy demand is highest in New York when the weather condition is snowy and partially cloudy and lowest when the weather is snowy, rainy, and foggy.



Although the energy demand in the state fluctuates when the weather conditions in the city change, a general trend is observed across all states; Texas constantly has the highest energy demand compared to other states, regardless of the weather conditions in Austin. New York has the lowest energy demand for all weather conditions in New York City compared to other cities and states.

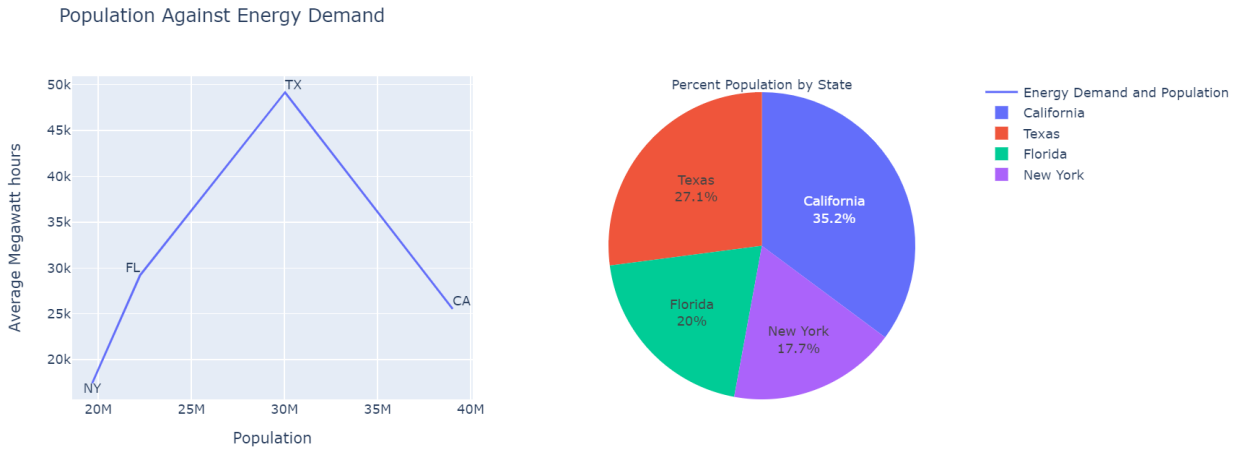
Total Energy throughout the Years



This line chart shows the yearly total energy demand in megawatt hours in Texas, California, Florida, and New York from 2016 to 2022. Texas has the highest energy demand, followed by Florida, then California, ending with New York. Texas and Florida's energy demand is approximately a hundred million megawatt hours apart. The difference between New York's and California's energy demand is close to a hundred-million-megawatt hour.

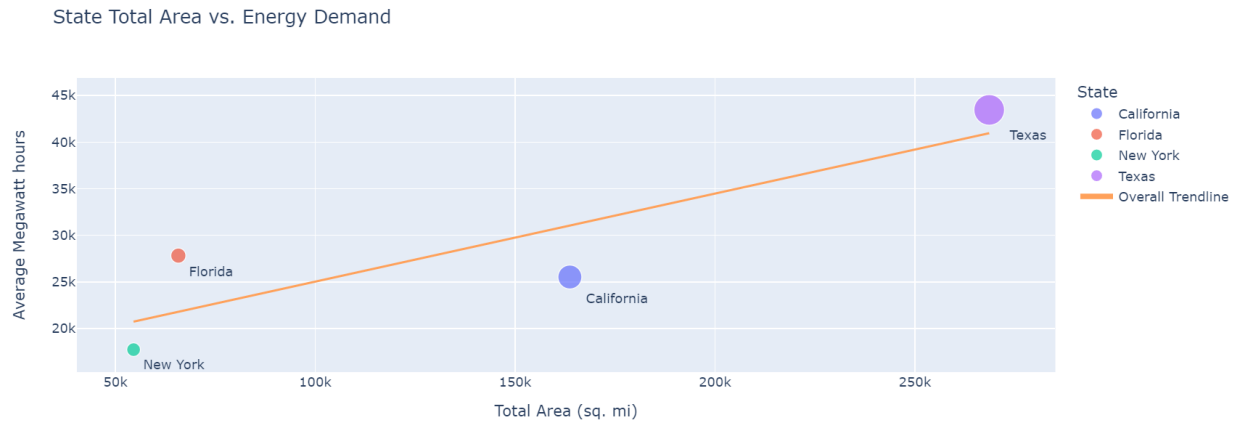
Energy Demand and Population

The scatter plot displays the hourly average energy demand in megawatt hours against the population in Texas, California, Florida, and New York. The pie chart represents the percentage population in each state, with California having the highest population, followed by Texas, Florida, and New York. California's energy demand is the second lowest on the scatter plot despite California having the highest population. Conversely, Texas has the highest energy demand while being the second most populous state.

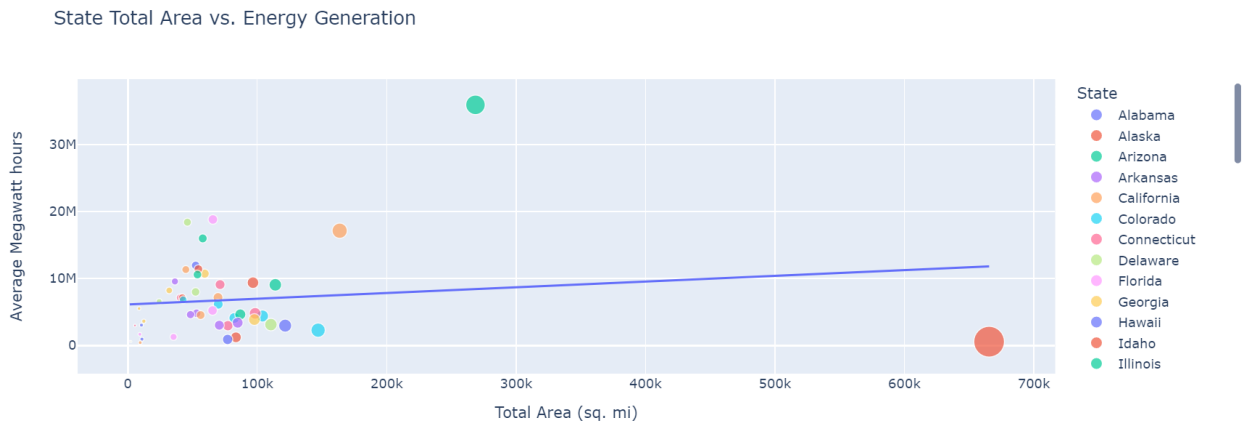


Energy Demand and State Size

This scatter chart displays the hourly average energy demand in megawatt hours against the total area for California, Florida, New York, and Texas. The trendline shows a strong positive correlation between state area and energy demand. As the state area increases, the energy demand increases.



However, the scatter chart represents data on four states and does not consider all 50 states. Therefore, examining beyond the scope of the four states' energy demand and state area is required. To further investigate energy demand against state areas in 50 states, a dataset on energy generation for all states was extracted from the website, Ember. It is worth noting that energy generation is on par with energy demand in the U.S. In other words, energy generation and demand are closely related. If there is a weak correlation between the state area and energy generation, there is also a weak correlation between the state area and energy demand.



This scatter chart shows the monthly average energy generation in megawatt hours against the total area for 50 states. Unlike the trendline in the previous scatter chart, this trendline does not show a strong correlation between state area and energy generation for all states.

Analysis of Texas 2021 Winter Storm

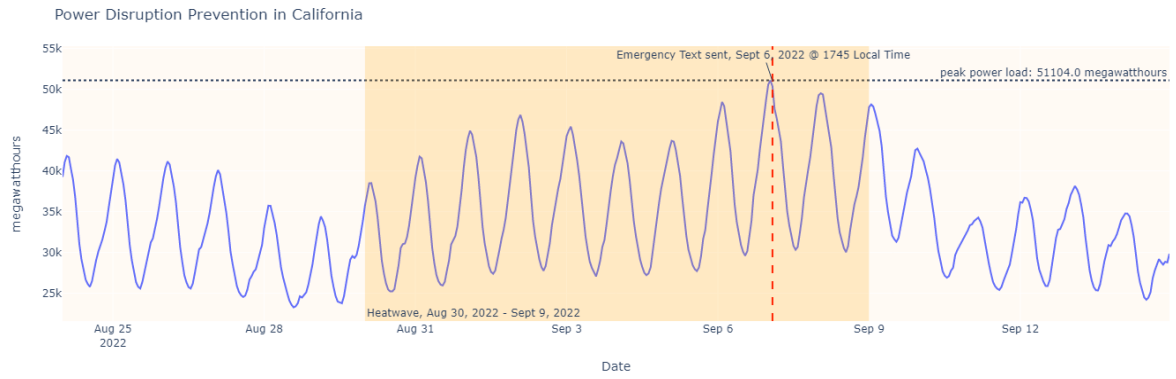
The line graph above depicts the rolling outages during the Texas winter storm of February 2021. Power demand peaked on the night of February 14th, as temperatures dropped below freezing and customers turned on their space heaters. Unable to meet the demand, coupled with power stations not being winter-proof, rolling outages began, cutting power to millions of Texans.

Analysis of Hurricane Irma in Florida

The line graph below shows power disruptions during Hurricane Irma in September 2017. Irma made landfall as a category-four hurricane with wind gusts of 130 mph. With power lines being disrupted due to the winds, approximately 64% of the population lost their power at the peak of the outages.

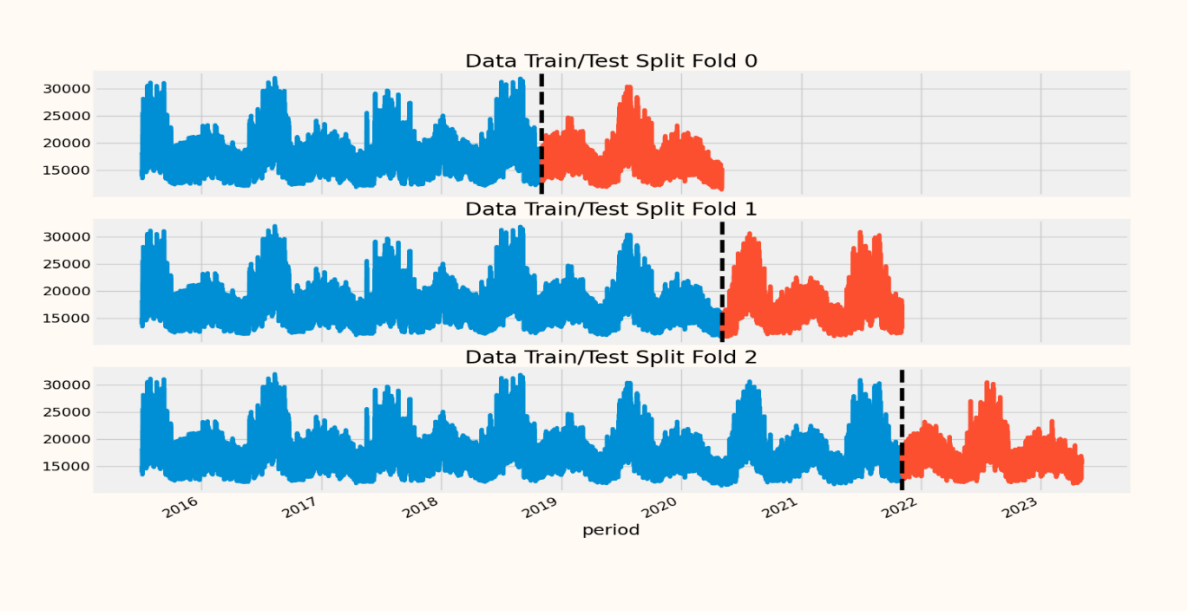
Analysis of California Heat Waves in 2022

This line graph shows the resilience of California's power grid during the heat wave in September of 2022. As temperatures rose, so did the power demand. But preparations such as increasing battery storage capacity and importing energy from neighboring providers prevented a widespread power outage for residents in California.



Machine Learning

We tested several ML models, including XGBoost, Random Forest, and Histogram Gradient Boosting but ultimately went with LightGBM. LGBM gave us speed and accuracy compared to the other models. For our training, we used cross-validation over three time periods. We accomplished this using SKlearns timeseriessplit. Below is a graph of how we split the data. We took the averages across the three folds as our model's final score. We scored using root mean squared error (RSME) and mean absolute error (MAE).



Our model tended to overfit, so we used hyperparameters to tune the model. For example, our unoptimized model for Texas began with an RSME of ~ 6000; after tuning, we got it down to ~3400. Below are the parameters we used and the final scores for each state.

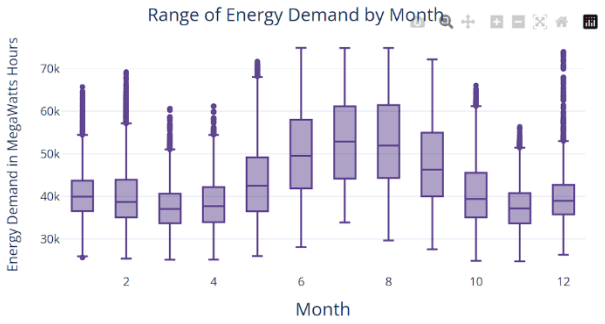
State	RMSE	MAE
California	1721.98	1200.40
Florida	2029.25	1464.15
New York	872.44	638.80
Texas	3484.53	2606.52

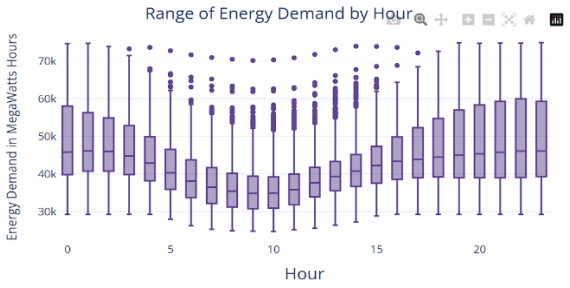
```
TX_params1 = {
    "n_jobs": -1,
    "boosting": "gbrt",
    "num_iterations": 10000,
    "early_stopping_round": 100,
    "max_depth": 10,
    "learning_rate": 0.005,
    "num_leaves": 30,
    "lambda_l1": 30,
    "lambda_l2": 0,
    "seed": 0,
    "metric": ['rmse','mae'],
    "device_type": 'gpu',
    "min_data_in_leaf": 12,
    "bagging_fraction": .8,
    "cat_smooth": 18,
    'max_bin': 80,
    "bagging_freq": 8,
    "min_sum_hessian_in_leaf" :250,
    'path_smooth':22,
    "feature_fraction": .6,
    "importance_type" :'split',
    "extra_trees":  False,
    'force_col_wise': True

}
```

Dashboard

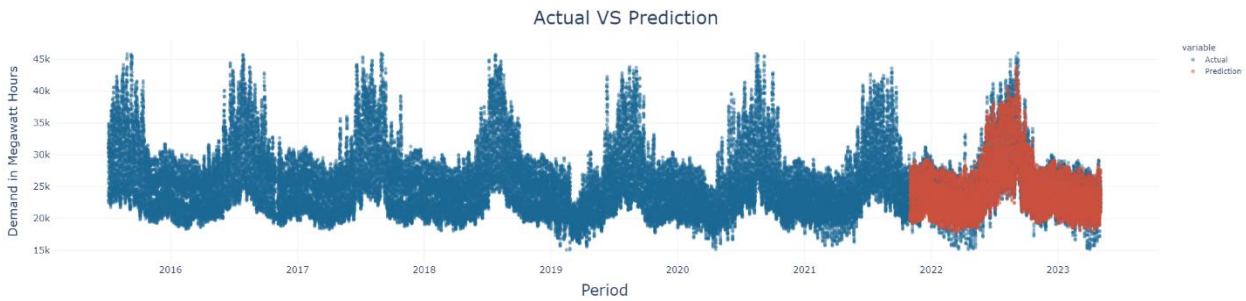
We constructed our dashboard using Dash and Plotly. It was then deployed and hosted on render. One similar thing across all four models is that peak demands are in the summer months and between 4 PM to 8 PM local time.



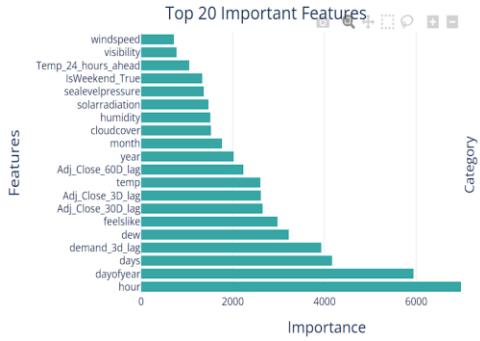


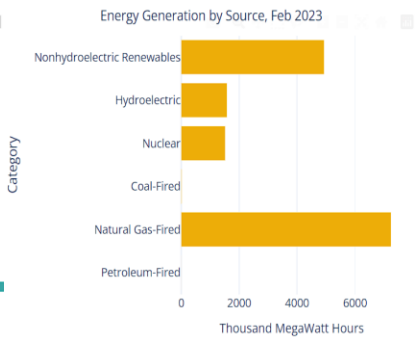
California Dashboard

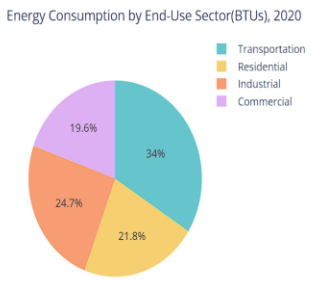
For California, we plotted the actual vs. prediction below, and we can see the model performs well in the mid-range but fails to predict the lowest and highest energy demands.



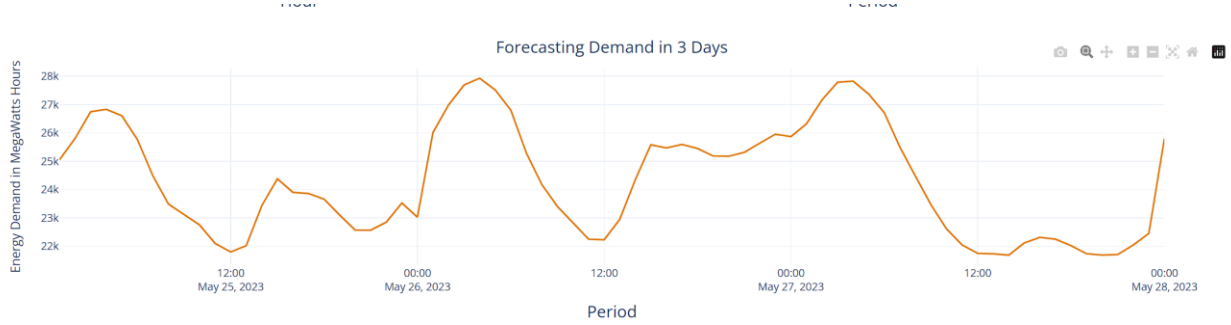
Looking at the important feature graph, we can see that the hour and day of the year are essential features for the California model. In addition, the adjusted close lags for XLE are higher in this model than in other models; this suggests that California may be more sensitive to energy price fluctuations. In the second graph, we can see that most of California's energy comes from natural and renewable gas. California has been heavily pushing to go fully renewable by 2035. Finally, the pie chart shows a closer distribution of energy consumption by each sector.





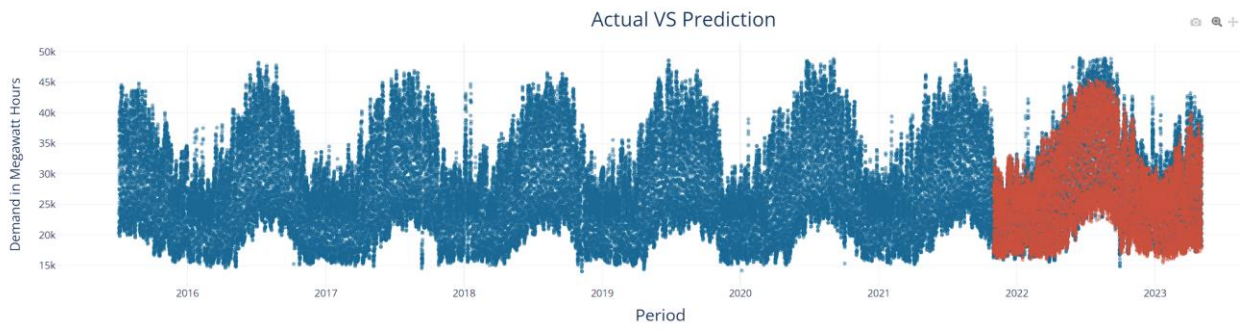


Our dashboard updates every 24 hours with the latest information pulled from the SQL server. Below is our model's forecast from May 25th, 2023, to May 28th, 2023.

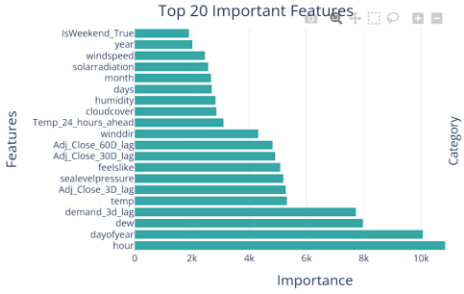


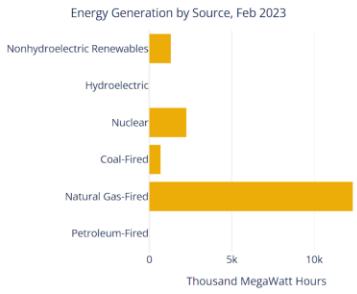
Florida Dashboard

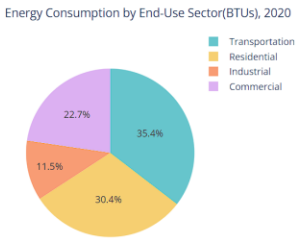
For Florida, our model could predict the low end very well but struggled to predict peak energy demand.



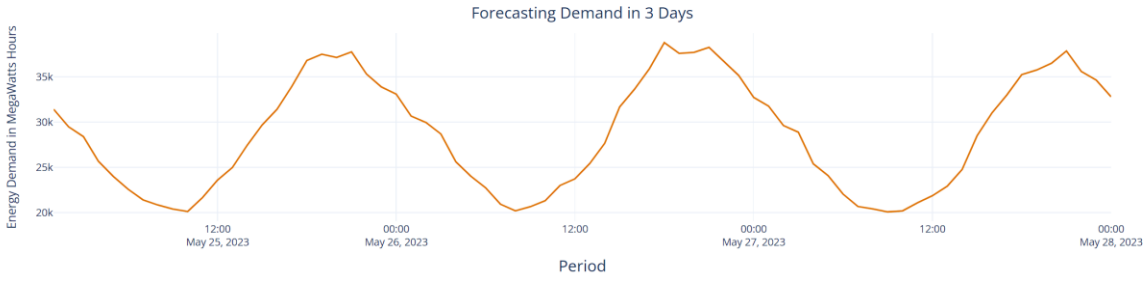
Our feature importance graph shows that sea level pressure is significantly larger than the other models. Sea level pressure across the different models is ~1300, while for Florida, it is 5194. After some research, a low sea level pressure usually indicates cloudy/stormy weather, while a high pressure indicates a clear sky. Florida does not have a diverse energy source. Most of its energy comes from natural gas.





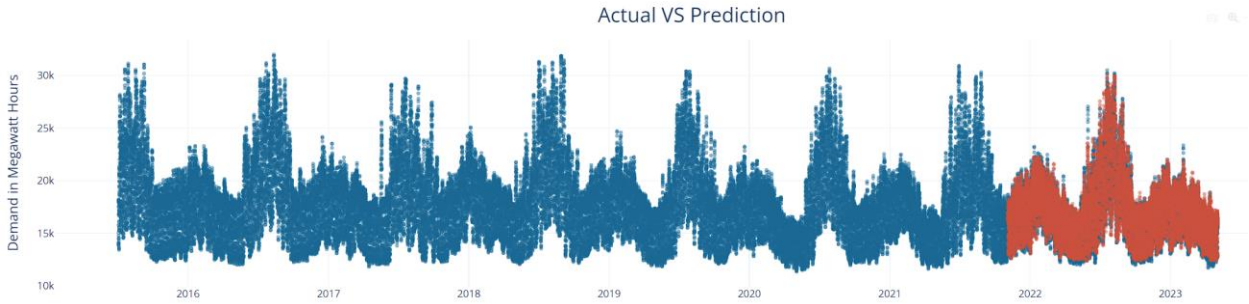


Finally, we have our forecast. Looking at historical data, we can see that energy usage is usually in 20-45K megawatt hour range during this time of year. Our forecast ranges from 20-35K.

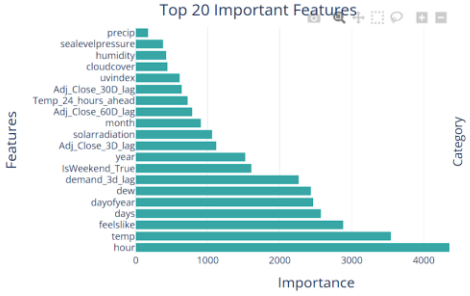


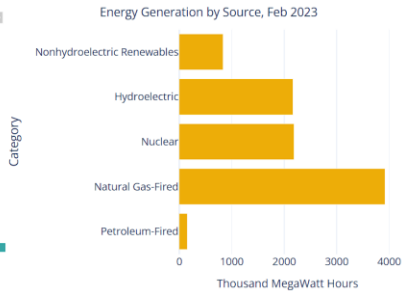
New York Dashboard

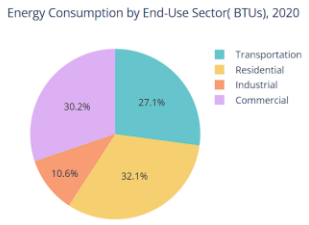
Our New York model performed the best with an RMSE of 872.44. The graph below shows that the N.Y. model captures both the low and high energy demand.



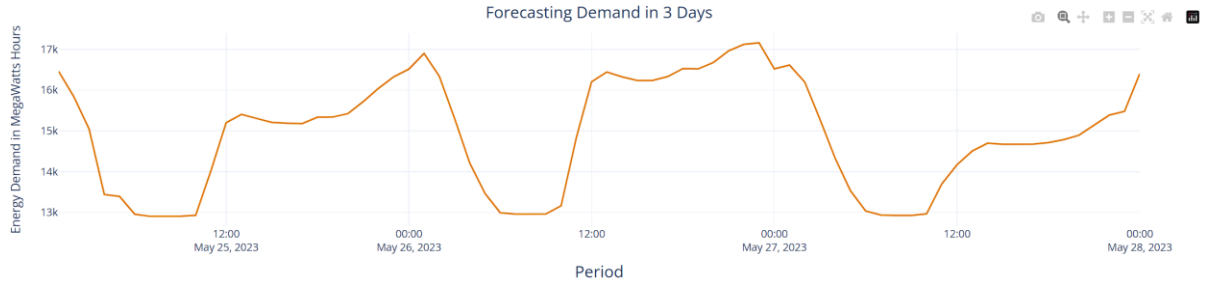
From the graphs below, we can see that N.Y. has a diverse energy source. They also have a small percentage of consumption by the industrial sector. The weekend seems to play a higher importance in N.Y. One explanation is that a third of N.Y.'s population is in New York City. It is possible that since NYC is a walkable city, a good portion of the residents are out and about during the weekends.





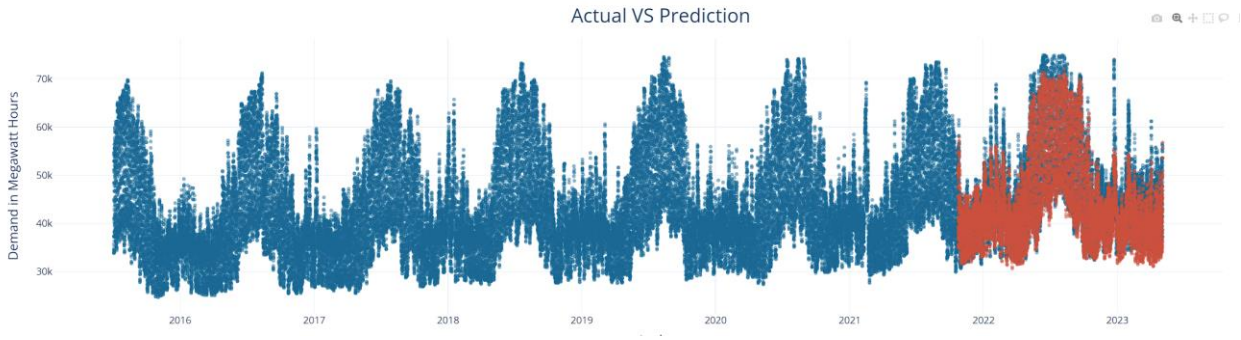


The graph below is our model's forecast for May 25th, 2023, to May 28th, 2023.

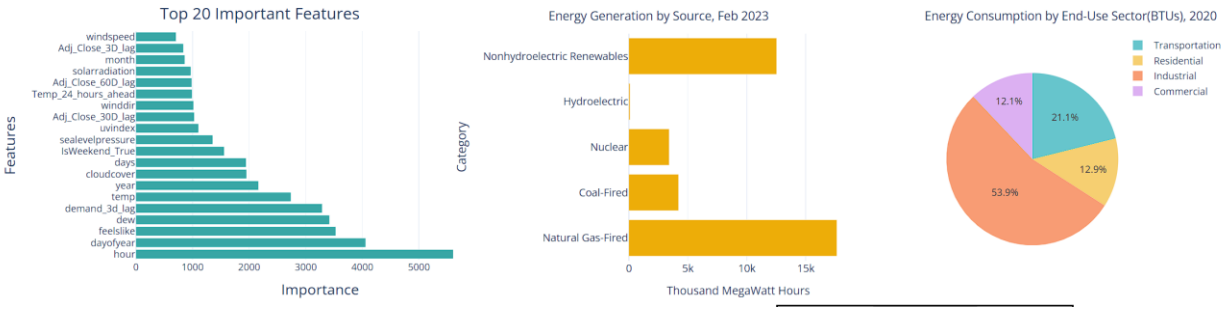


Texas Dashboard

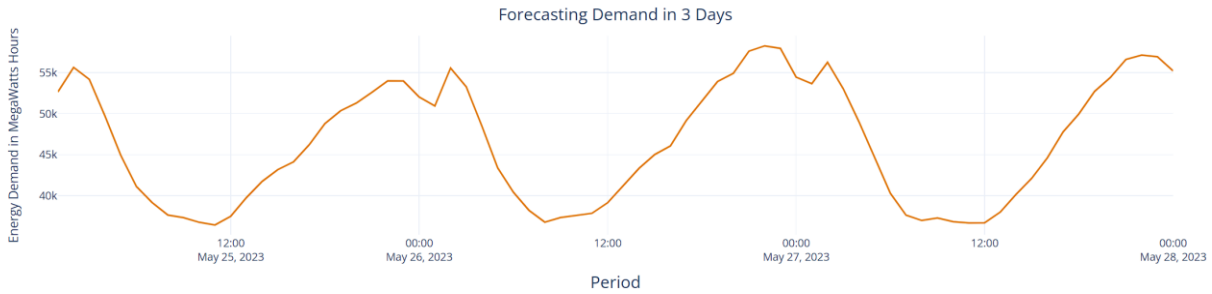
Texas is our worst-performing model, with an RMSE of 3484.53. Texas seems to have days where energy demand spikes but those that are frequent enough not to be outliers. Our model struggles to predict these spikes in energy demand.



The pie chart below shows one theory we have for these spikes. Most of Texas's power consumption comes from the industrial sector. The industrial sector usually consists of factories and warehouses. Factories are not continuously operating at 100% capacity; they are scaled up and down based on demand. The high level of consumption by the industrial sector adds to the variance in the model, which our model does not have information on. It would be interesting to find additional data on factories in Texas that we could add to our model to see if we can improve its importance.



Finally, below is our forecasted data for Texas. Our model predicts a peak usage of 58K megawatt hours on May 26th, 2023.



Conclusion

We observed a correlation between weather conditions and energy demand. When the weather conditions change, so does the energy demand. However, the fluctuations in state energy demand based on city weather conditions were not drastic enough to change the overall trend of Texas requiring the most energy, followed by Florida, California, and then New York. Additionally, the state population size does not affect the state's energy demand. Furthermore, while there is a positive correlation between state area and energy demand in just the four states, there is no strong correlation between state area and energy generation in all 50 states.

References

CAISO. (2022, November 2nd). California ISO posts analysis of September heat wave. https://www.caiso.com/Documents/california-iso-posts-analysis-of-september-heat-wave.pdf

Director, L. M. E., Metzger, L., & Director, E. (2022, December 28th). *The Texas freeze: Timeline of events*. Environment Texas Research & Policy Center. https://environmentamerica.org/texas/center/articles/the-texas-freeze-timeline-of-events/

Fulghum, N. (2023, April 26). *U.S. electricity data (state-level)*. Ember. https://ember-climate.org/data-catalogue/us-electricity-data/

Hodge, T., & Lee, A. (2017, September 20th). *Hurricane Irma cut power to nearly two-thirds of Florida's electricity customers*. Homepage - U.S. Energy Information Administration (EIA). https://www.eia.gov/todayinenergy/detail.php?id=32992

The Fact File (2022, January 25th). *50 States Ranked By Size, In Square Miles*. https://thefactfile.org/50-states-area/

U.S. Census Bureau. *State Population Totals: 2010-2019*. https://www.census.gov/data/datasets/time-series/demo/popest/2010s-state-total.html#par_textimage_1873399417

U.S. Census Bureau. *State Population Totals and Components of Change: 2020-2022*. https://www.census.gov/data/datasets/time-series/demo/popest/2020s-state-total.html

U.S. Department of Commerce, N. (2018, July 23rd). *Hurricane Irma Local Report/Summary*. National Weather Service. https://www.weather.gov/mfl/hurricaneirma